

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
N.V. PHILIPS' GLOEILAMPENFABRIEKEN

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, HOLLAND

X-RAY TUBES WITH ROTATING ANODE ("ROTALIX" TUBES)

by J. A. van der TUUK.

621.386.1

X-ray tubes with a rotating anode permit a much higher specific loading of the focus than tubes with a stationary anode, a property which is of great importance for the röntgenography of moving organs, such as the heart and lungs. While, however, for photographs of heart and lungs only low tube voltages (< 70 kV) and short loading times are required, for other diagnostical applications (especially for examinations of the stomach) the X-ray tube with rotating anode must also be able to be operated at very high tube voltages (up to 100 kV) and be loaded continuously with 150 W. A type of tube with a rotating anode is here described, which satisfies all these requirements very well. Due to the fact that the tube, which is entirely of hard glass, is immersed in oil, its dimensions and weight could be reduced considerably compared with previous constructions. The anode consists of a disc of solid tungsten, while in the original „Rotalix” tubes the anode was a block of copper with a thin layer of tungsten on its front surface. The detailed discussion of the advantages and disadvantages of the two constructions presents an opportunity of dealing with a number of typical problems connected with the rotating anode.

When in 1929 there first appeared on the market a practically useful X-ray tube with rotating anode — the “Rotalix” tube developed in the Philips laboratory¹⁾ — it meant a big step ahead in the quality of the X-ray photographs which could be made in the medical diagnosis of moving organs such as the heart and lungs. In the case of moving objects short exposure times are of course desired. Thus for a given blackening of the film a high X-ray intensity is required, and that requires a high specific loading of the focus of the X-ray tube, since that focus must be kept small for the sake of a small half shadow (geometrical lack of sharpness) in the X-ray picture. The specific focus loading is limited by the temperature of the focus, which rises during the time of loading and which may not exceed a certain value because of the evaporation of the anode material. Because of the fact, that, in the case of the rotating anode, each point on the surface of the anode serves as focus only for a small part of the total exposure time, the specific loading of the focus can be made considerably higher than in the case of a stationary anode. With exposure times of less than about 0.04 sec., as was discussed previously²⁾, the theoretically attain-

able improvement in the specific loading amounts to a factor

$$p = \sqrt{2\pi n r t / f} \dots \dots \dots (1)$$

r is here the radius of the path of the middle of the focus, n the number of revolutions, t the time of exposure and f the width of the focus. The practically attainable improvement factor is somewhat smaller for various reasons, than follows from the formula: with a “Rotalix” tube of the original construction, with $r = 2$ cm, $n = 2900$ r.p.m. and $f = 2$ mm, for an exposure time of 0.04 sec., $p = 8$ to 9.

The result of this improvement in the specific focus loading in the case of photography of moving organs meant approximately a reduction by one half of the lack of definition of the X-ray picture (with the same contrast), since this lack of definition is practically³⁾ proportional to $\sqrt[3]{p}$, when one works under the optimum exposure conditions. Diagnos-

²⁾ See: J. H. van der Tuuk, The “Rotalix” Tube for X-ray Diagnosis, Philips techn. Rev. 3, 272, 1938. The arguments given here very briefly will be found in more detail in that article.

³⁾ See for example the explanation of the optimum exposure conditions in: A. Bouwers and G. C. E. Burger, X-ray Photography with the Camera, Philips techn. Rev. 5, 255, 1940.

tically this improvement in definition was so valuable that the photographs obtained with a tube provided with a rotating anode could be said to be of a higher class.

It would appear from formula (1) that the factor p can be made still larger, for instance by giving the anode a larger diameter and making it rotate more rapidly. However, the additional improvement, which can practically be obtained in this way is relatively small compared with that already achieved, since p increases only proportionally with $\sqrt{rn}$. A large increase in the specific loadability is not readily possible, while the necessity of a slight increase cannot at present be considered to be actual, since in order to guarantee optimum exposure conditions the photographic industry would have to increase considerably the sensitivity and the fineness of grain, respectively of the combination X-ray film plus intensifying screen³). It is thus

tively low (45 to 70 kV) and the tube is always loaded only for a short time. It soon became evident, however, that the tube with rotating anode should also be able to be used for other diagnostical purposes, especially for stomach exposures and for fluoroscopy, since doctors usually prefer to carry out all those examinations with the same tube⁵). The tube therefore had to be made suitable also for the high voltages (up to 100 kV for example or even higher) necessary for stomach photography and for relatively heavy continuous loading (100 W and sometimes even 200 W), while in connection with easy adjustability for the different kinds of exposure, the dimensions and the weight of the tube had to be limited as far as possible.

In this article we shall give a description of a new „Rotalix” tube developed in recent years. It will be shown how and to what extent the above-mentioned requirements are satisfied.

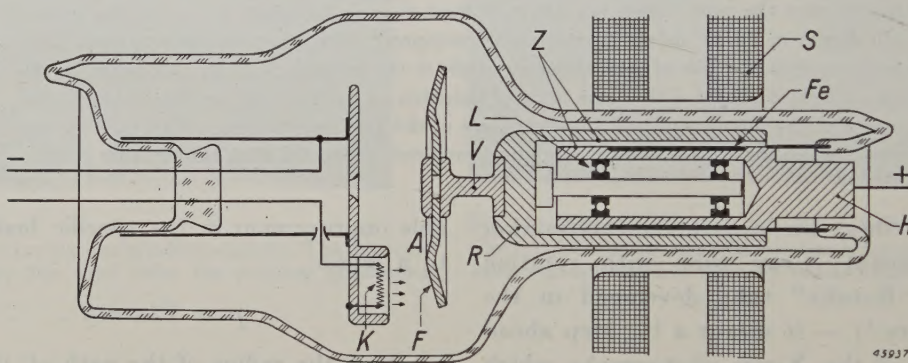


Fig. 1. Diagram of the cross section of the new „Rotalix” tube. K filament, F focus on the anode A which consists of a solid tungsten disc. The latter is fastened by means of a molybdenum connecting piece V to the copper rotor R which turns in two ball bearings L in the anode support H and is driven by the stator S . Fe is an iron ring for the purpose of concentrating the lines of force in the rotor: Z blackened surfaces for better radiating the developed heat. The anode support and the ball bearings remain practically cold, which is essential for a good working of the bearings. Behind the cathode a barium getter is applied.

understandable that the different types of tube with rotating anode which have been developed since the beginning⁴) do not aim primarily at a further increase in the specific loadability of the focus. The new constructions are characterised rather by a tendency towards greater simplicity and security of operation and towards an extension of the possibilities of application. As concerns the latter, it may be noted that the „Rotalix” tube was originally intended for the photography of heart and lungs, where it is mainly a question of a high specific focus loading, while the tube voltages may be rela-

Main characteristics of the new „Rotalix” tube

Fig. 1 shows a simplified cross section of the tube. The anode in this case consists of a disc of solid tungsten fastened by means of a short molybdenum connecting piece to a copper rotor. This part of the tube is surrounded externally by a stator fed with polyphase current which furnishes the motive power. In fig. 1 may also be seen the filament excentrically opposite the outer edge of the rotating anode disc. The (primary) electrons emit-

⁴) See for example A. Ungelenk, Fortschr. Röntgenstr. 49, 162, 1934.

⁵) In stomach examinations it is essential that the tube can also be used for fluoroscopy. Fluoroscopy and photography here follow one another immediately, as will be discussed later.

ted by the filament are focussed on the anode. The tube wall is made of a special kind of hard glass which is very resistant to the bombardment by the secondary electrons which are emitted by the anode⁶⁾. The whole tube is housed in an earthed metal jacket which protects the operator from contact with parts carrying high voltage and helps to cut off the X-rays in undesired directions. (see fig. 2 and 3). The jacket is filled with oil.

When we compare this construction with that

and the earthed housing. While the older tube with housing was 83 cm long and had a maximum width of 22 cm, in the tube according to fig. 2 these dimensions could be reduced to 46 en 14 cm, respectively. The new tubes are adopted in all respects to voltages of up to 100 kV. It is also of importance that the breakdown resistance is ensured independent of atmospheric conditions. Tubes in air, which are intended for use near sea level, cannot in general be loaded with the highest voltages in

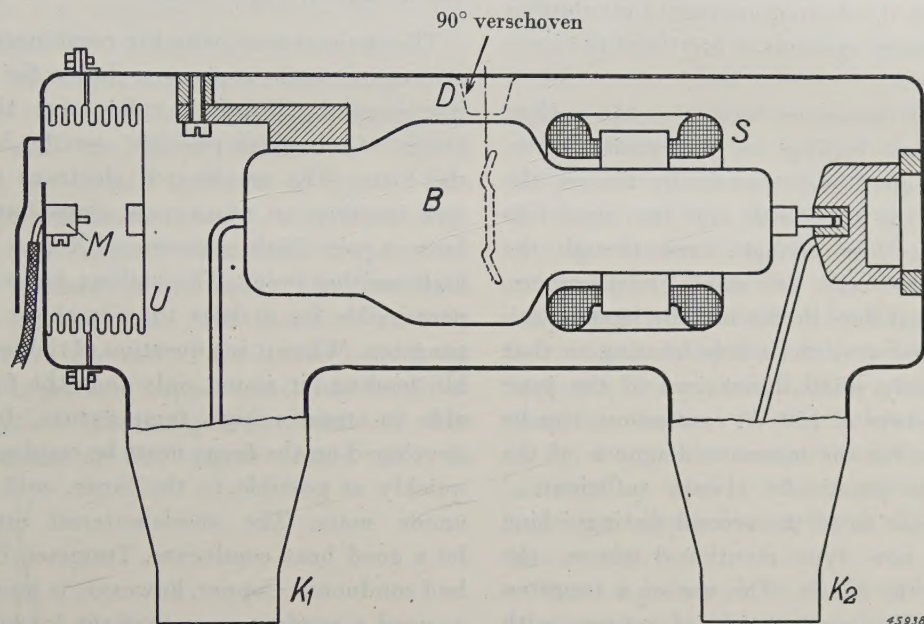


Fig. 2. The *B* tube of fig. 1 mounted in the metal housing *S* is the stator. At *K*₁ and *K*₂ are the connections for the high-voltage cables. *U* accordion shaped box for sealing off the oil-filled space. Because of this shape the oil can expand when warm without appreciably increasing the pressure. If the temperature of the oil rises above a certain value, the compression of *U* closes a switch *M* which switches off the supply voltage of the X-ray tube. *D* is a cup made of insulation material (shifted about 90° in position in the drawing) which keeps the oil away from the spot where the X-rays leave the tube. In this way the absorption of the useful X-rays by the oil is kept as small as possible. By making the tube wall and the thickness of *D* small, the filter of the tube with housing is kept small, namely less than corresponds to a thickness of 1 mm of aluminium. The housing is made of iron with a thin lead covering on the inside to prevent the passage of X-rays.

of the older “Rotalix” tubes, which have already been described²⁾, there are two obvious differences. The older tubes were surrounded by air instead of by oil; and the anode had an entirely different construction. It consisted of a heavy copper block upon the front of which was fixed a thin layer of tungsten (lozenge) and whose rear side was provided with special cooling jackets.

The immersion in oil has certain important advantages⁶⁾. Due to the high break down resistance of the oil it was possible to use relatively small insulation distances between the poles and between the latter

the mountains (sanatoria), since at the lower atmospheric pressure the breakdown resistance of air is lower. In spite of the oil filling the total weight (17.5 kg.) of the new tube with housing is still appreciably lower than that of the tube with air insulation.

The insulation by means of oil also has advantages in the driving of the rotating anode. In order to use as little power as possible for that purpose, the „air slit” between the stator and the rotor at high voltage should be as narrow as possible. A possible solution⁴⁾, and the one used in the case of air insulation, is to have the stator with its windings at the same high potential as the anode. A special transformer is then necessary for the feeding; one

⁶⁾ Cf. also: J. H. van der Tuuk, Hard glass X-ray Tubes in Oil, Philips techn. Rev. 6, 309, 1941.

with a secondary winding which is insulated for high voltage. Placing the tube in oil, however, makes this complication entirely unnecessary, since even with the stator earthed, the air slit can be made very narrow with no danger of flashover from the glass wall, so that only 500 W during 0,5 sec. are necessary to start the rotor. In the case of the previously described „Rotalix” tube, thanks to the practically earthed metal intermediate section, an „earthed” stator could also be employed, but in that case about 1100 W during 0,8 sec. were necessary just to reach the desired number of revolutions (because of a greater moment of inertia of the rotor and a wider air slit).

Finally the oil immersion facilitates the cooling during continuous loading to no small degree. Because of the good heat transfer by the oil, the temperature of the tube wall and the insulation material remains low enough, even though the dimensions of the tube are small. Furthermore, due to the natural flow in the oil, the heat is uniformly distributed over the whole housing so that with the relatively small dimensions of the type shown fig. 2 a power of 150 W, continuous, can be dissipated. Even for the intensive diagnosis of the stomach, this is practically always sufficient.

Let us now pass on to the second distinguishing feature of the new type mentioned above, the construction of the anode. The use of a tungsten disc instead of the combination of copper, with a thin lozenge of tungsten means a simplification in various respects. However, a comparison be-

tween the two types of anode is not complete, simply by mentioning this fact. The transition to the construction now employed involved a give and take, and it is quite comprehensible that the older construction will continue to be preferable for certain special diagnostical purposes. It seems worth while to go somewhat more deeply into the problems here concerned, especially because the specific problems of the rotating anode are thereby strikingly emphasised.

The tungsten-copper anode

The main reason, why the combination described of tungsten and copper was chosen for the anode of the original „Rotalix” tubes, was the desire to attain the highest possible specific loadability of the focus. The accelerated electrons must in any case impinge on tungsten, since that metal combines a very high emission of X-rays with a very high melting point. The highest focus temperature permissible for a short time is about 2500 °C for tungsten. When it is a question of the heaviest possible loading, it is not only that the focus must be able to stand a high temperature, but the heat developed on the focus must be conducted away, as quickly as possible, to the inner, cold parts of the anode mass. The anodematerial must therefore be a good heat conductor. Tungsten itself is not a bad conductor. Copper, however, is more than twice as good a conductor, so that the loadability will be increased when the anode material, as close as possible, behind the focus is not of tungsten but of copper.

Closer consideration shows that this reasoning is still somewhat too simple. The tungsten layer cannot be made arbitrarily thin, since the copper directly behind it must not become hotter than a maximum of 1000 °C (meltingpoint 1080 °C). The minimum required thickness of tungsten can be determined by considering the variation of temperature in a longitudinal cross section of the anode, when the extremity of the cross section lying in the front surface of the anode ($x = 0$) is exposed to the beam of electrons from the moment $t = 0$ to $t = t_1$. Here t_1 is the time, during which the anode rotates one width of focus farther; thus with the above mentioned values of r , n and f , t_1 is 0.00033 sec. At each point the temperature will first rise and then, after the loading has ceased, it will fall again as the heat is conducted farther away toward the inside of the anode. If we now plot the maximum increase in temperature T , which is reached in every separate point, we obtain, with a solid tungsten anode, a curve like that reproduced in fig 4a. The load is there assumed to be such,

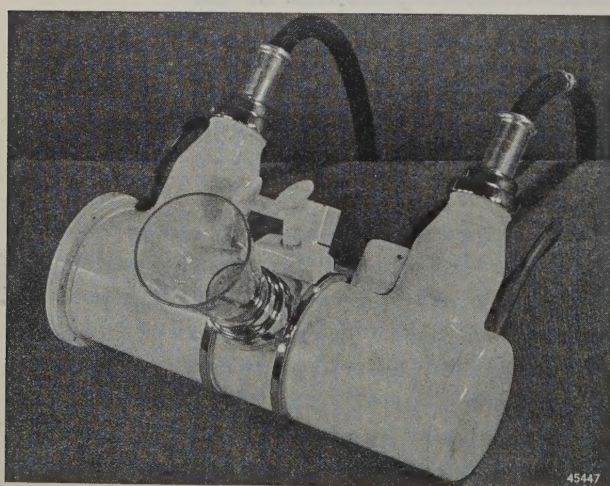


Fig. 3. The new „Rotalix”-tube in its housing, the whole having a length of 46 cm. The housing is in a clamping-strip turnable around its axis. The leadglasstube, through which the X-rays pass, is visible. The housing is of special finish with built in cooling spiral, by which the power could be increased to 600 W. Besides the contacts for the high tension cables rubber tubes can be seen for the leading in and out of the cooling water. The normal construction, intended for 150 W, does not possess a cooling spiral.

that the maximum temperature increase of the focus itself ($x = 0$) which occurs at the moment $t = t_1$, possesses exactly the highest permissible value. It is now clear, that the anode can only consist of copper beginning at a certain depth x_1 , for which the maximum temperature ever occurring is not more than 1000°C . Since, as will appear later, we must count on a basis temperature of for example 450°C for the copper, we can find x_1 , in the figure as the point where the curve *a* has fallen to $T = 550^\circ\text{C}$. In that way we find that, according to this curve, the tungsten layer must extend at least to a depth of 0.2 mm. If we now draw the temperature distribution in the longitudinal cross section at the moment t , at which the loading is interrupted and the focus reaches its highest temperature, we obtain curve *b* of fig. 4. This shows that, at that moment only, very little (5%) of the total heat applied penetrated deeper than 0.2 mm into the body of the anode. The maximum temperature of the focus will therefore scarcely be affected, when the rest of the anode behind the first 0.2 mm is made of copper instead of tungsten.

We thus reach the unexpected conclusion that the employment of the copper would offer no

advantage with respect to loadability. This is, however, true only when the exposure time is so short that the anode does not make more than one complete revolution during the loading, thus for exposures of less than $t_2 = 0.02$ sec. If the loading lasts longer each point on the surface of the anode is struck not once, but repeatedly by the beam of electrons. Since in the time interval between the first and the second loading the temperature of each point in the orbit of the focus does not decrease from the maximum $T_1 = 2500^\circ\text{C}$ to the initial temperature $T_0 = 450^\circ\text{C}$, but only to a certain value $T_2 > T_0$, the second loading must be lower, the third still lower, etc. The advantage of the copper now becomes evident. In the time t_2 , which, with the above-mentioned value of r and f , is about 60 times as long as t_1 , the greater part of the heat ($>80\%$) penetrates deeper than 0.2 mm into the anode. The temperature increase $T_2 - T_0$ remaining after the first revolution is thus practically entirely determined by the properties of the more deeply lying material, and is therefore, when copper is chosen, only half as great as with a solid tungsten anode, namely about 60° . Curves *c* in fig. 4 give the temperature distribution at the moment t_2 : with solid tungsten (fullline curve) and with copper from a depth of 0.5 mm (broken-line curve).

With exposure times of the order of 0.05 to for instance 3 sec., the copper in this way gives an appreciable improvement in the permissible loading. Much longer exposure times scarcely occur in the use of tubes with rotating anodes in ordinary X-ray diagnostics. It may, however, happen that a large number of successive exposures of a moving object must be taken at short intervals (röntgen cinematography), which amounts to total loading times of for example 20 sec. or more. With such long times, the heat developed on the front surface of the anode penetrates not only to the inside, but even to the rear side of the anode, in other words, the necessarily limited dimensions and the consequently limited heat capacity of the anode now begin to make themselves felt. The residual temperature of the focus, which steadily increases with increasing loading time, will therefore increase even more rapidly and the permissible loading will decrease even more rapidly. If, in order to be able to use the tube for cinematography, it is desired to make the heat capacity of the anode fairly large, it is therefore fundamentally of advantage to use copper. Copper is much cheaper than tungsten and, for the same heat capacity, only $3/4$ of the volume and $1/3$ of the weight of copper is needed.

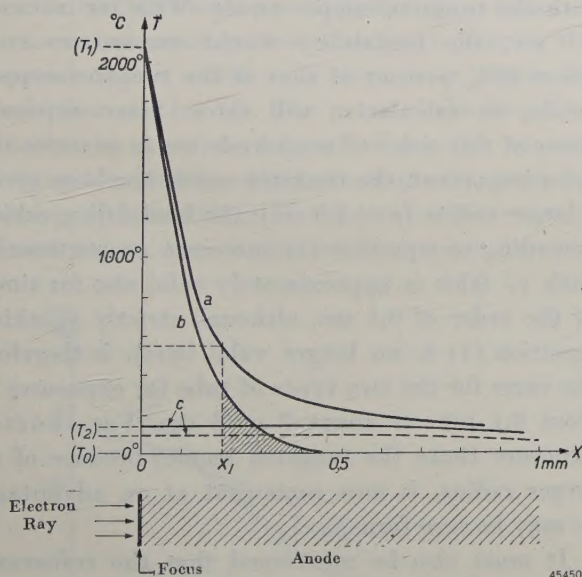


Fig. 4. Temperature distribution over a longitudinal cross section of the anode. The increase of temperature T above the basis temperature (T_0) of about 450°C is plotted as a function of the depth x beneath the focus. *a*) Maximum temperature ever occurring during or after the loading; anode of solid tungsten, loading from $t = 0$ to $t_1 = 0.00033$ sec. (i.e. the time during which the focus is displaced by its own width), such that the focus ($x = 0$) takes on the highest permissible temperature, ($T_1 = 2500^\circ\text{C}$). *b*) Temperature at the moment t_1 , at which the loading ceases. *c*) Temperature at the moment $t_2 = 0.02$ sec. i.e. after one complete revolution; full-drawn line: anode of solid tungsten, broken line: anode of copper from a depth of 0.5 mm. (All curves borrowed from W. J. Oosterkamp, Thesis, Delft 1939).

7) Those considerations are borrowed from: W. J. Oosterkamp, thesis Delft, 1939.

If we go a step farther with the loading time we enter the region of fluoroscopy, where we may speak of continuous loading. The power which can here be applied is no longer determined by the focus temperature but by the temperature taken on by the anode as a whole⁸). It is now found that, when copper is used, there are definite limitations: for a copper anode covered with tungsten the permissible temperature lies much lower than for a solid tungsten anode. While for the copper immediately behind the lozenge it amounted to about 1000 °C for a very short exposure in the case of continuous loading of the whole anode mass a temperature of not higher than about 450 °C (the basis temperature assumed above) is permissible, since at higher temperatures all kinds of undesired things take place: recrystallisation phenomena occur in the copper, the mechanical strength decreases, and the requirements are fairly high in that respect because of the rapid rotation, too much vaporized copper enters the tube (the vapour pressure of copper at 850 °C is already 10^{-5} mm; the vapour pressure of the copper oxide from various sources is even noticeable at 500 °C) and finally traces of residual gases may be freed from the anode. The latter may be ascribed to the fact that, during the evacuation and outgassing of the tube, the temperature may not be raised too much because of this very restriction in the permissible temperature of the copper. Now since the dissipation of heat by the anode, which rotates in a vacuum, can only take place by radiation, and since the radiation increases with the fourth or fifth power of the temperature, the restriction of the basis temperature of the anode to 450 °C makes it difficult to reach an adequate power for fluoroscopy. Only by employing the fairly complicated construction, which has been described in the articles already cited²), and which consists in an artificial enlargement of the radiating surface of the anode and an improvement of its emissivity, was it possible to increase the heat dissipation and thereby the fluoroscopic power to reasonable value.

The combination of tungsten with copper was accompanied by another complication. The brief, very high specific loadings result in corresponding high temperature gradients and mechanical tensions in the rotating anode. These make necessary a special solution of the problem of joining the tungsten lozenge to the copper (as was also discussed in the article cited²) a solution which made

very heavy demands on construction and fabrication.

The tungsten anode

The simplifications resulting from the use of tungsten alone for the anode are obvious. No precautions are necessary to keep the basis temperature of the anode low with continuous loading. The anode can be rigorously outgassed during fabrication, which promotes particularly the safety from high voltage of the X-ray tube⁹). The problem of attaching the tungsten lozenge no longer exists.

From the foregoing it will have become clear that certain sacrifices must be made for these simplifications. We shall discuss these in somewhat more detail.

With exposure times shorter than the period of one revolution of the anode (< 0.02 sec.) the loadability of a tungsten anode, as we have seen above, will not be appreciably smaller than that of a tungsten-copper anode of the same radius and speed of revolution. With longer exposure times where the permissible loading falls as a result of the residual temperature upon repeated loading of the focus, the tungsten anode cannot compare with the tungsten copper anode. With for instance 0.1 sec. the loadability would amount to only about 80% percent of that of the tungsten-copper anode, as calculation will show. Since exposure times of this order of magnitude are in practice the most important, the tungsten anode has been given a larger radius ($r = 3.0$ cm); the loadability, which according to equation (1) increases proportionally with $\sqrt{r}$ (this is approximately valid also for times of the order of 0.1 sec. although strictly speaking equation (1) is no longer valid here), is therefore the same for the two types of tube for exposures of from 0.1 sec. to about 2 or 3 sec. For shorter exposure times the tungsten anode, because of its larger radius, is now somewhat at an advantage, as may be seen from fig. 5.

It must also be mentioned that the transverse dimensions of the tube with housing, inspite of the larger diameter of the anode in the new „Rotalix” tube, are not larger, but are even smaller than those of the old tube, thanks to the employment of oil insulation.

When we consider the further course of the curves in fig. 5 we see that the tungsten anode, which is now equal to the tungsten copper anode (of smaller diameter) from about 0.1 sec. to about 3 sec. is

⁸) See the communication of G. C. E. Burger, Hand. Ned. Nat. Gen. Congr. 1943, p. 286.

⁹) It may be mentioned that this has aided to make possible the driving of the „Rotalix”-tube on A.C. We hope to return on this question at another opportunity.

much inferior at still longer exposure times. This is due to the much lower heat capacity of the tungsten anode. The tungsten copper anode of the old “Rotalix” has a weight of more than 1 kg. for the same heat capacity the tungsten anode would, as stated above, need to be three times as heavy. Apart from the fact that the employment of 3 kg. of tungsten is unnecessary for our purpose and would accord very ill with the aspiration toward limitation of expense and weight, the present position of technology would make it practically impossible to make a block of tungsten of that weight with the desired good mechanical properties. The high melting point which among other properties renders tungsten so particularly suitable as material for anti cathodes, however, also makes it impossible to prepare massive tungsten objects by the ordinary metallurgical processes, namely fusing and casting. The methods of powder metallurgy must here be applied¹⁰). Pure tungsten in powder form, prepared chemically, is pressed into a rod which is then heated to a high temperature for some time (presintered and sintered). After this sintering, which is accomplished by heating with an electric current conducted by the rod, the ductility necessary for further working and application is obtained by means of a mechanical treatment. Finally the material can be rolled to a sheet from which the anode disc can be cut. The mass of the disc thus obtained is therefore determined by the amount of tungsten powder which is originally used for pressing and sintering. The amount which can be treated at one time is limited for various reasons,

for example because of the fact that with increasing mass the very brittle presintered rods become increasingly difficult to handle and further the fact that the currents required for sintering become extremely large, etc. In the methods of fabrication employed by us the finished tungsten discs have a weight of about 225 g. We shall return to this subject later.

The heat capacity of the tungsten anode is therefore only 1/12 of that of the copper anode. On the other hand, however, the temperature of the tungsten disc may be much higher than that of the copper, but this advantage does not compensate for the much smaller heat capacity. In this way therefore the loadability of the tungsten anode for very long exposure times (röntgencinematography) is much lower than that of the tungsten-copper anode.

Although the small mass of the tungsten anode is indubitably a disadvantage in this respect it may not be forgotten that in all kinds of other respects it is an advantage, particularly in connection with the bearings and driving of the anode. Since the tungsten anode together with the copper rotor only weighs about half as much as the tungsten-copper anode of the old tubes, it is easier to limit sufficiently the friction and wear on the ball bearings on which the anode turns¹¹). Furthermore, since the moment of inertia of the anode is only about 1/3 of that of the tungsten-copper anode it reaches the required speed of revolution with the same driving power much more quickly after being started. This is important for stomach examinations where the doctor uses fluoroscopy and requires immediately to fix the picture photographically at a definite moment. Since fluoroscopy takes place as a rule with stationary anode or at least with the anode not rotating at full speed, the necessary delay in photographing is determined by the time necessary for the anode to obtain full speed and for the heating of the filament which operates at a lower temperature for fluoroscopy (lower tube current of for instance 2 to 3 mA) than for photography. Reaching full speed from a state of rest, which took 0.8 sec. for the tungsten-copper anode, easily takes place in less than 0.5 sec. with the tungsten anode, thanks to the smaller moment of inertia. In order to profit by this, of course, the time necessary for

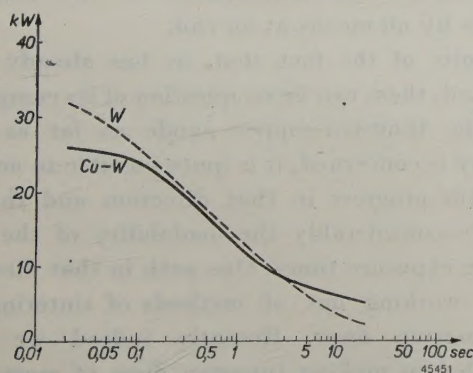


Fig. 5. Loadability (in kW) as a function of the loading time for a tungsten-copper anode (full-drawn line) and a tungsten anode (broken line). Loaded with pulsating direct current obtained with four values. The diameter of the anode at the middle of the focus is 4 cm with the tungsten-copper anode and 7 cm with the tungsten anode. The basis temperature is 450° C in the first case and 800 °C in the second. The speed of revolution is 2900 r.p. m. in both cases, the width of focus 2 mm.

¹⁰) See: J. D. Fast, The preparation of Metals in a Compact Form by Pressing and Sintering, Philips techn. Rev. 4, 309, 1939.

¹¹) The ball bearings are lubricated with a soft metal, in the Philips tubes lead is used. The vapour pressure of lead is extremely low so that the high vacuum in the tube remains so complete that the extremely high requirements of reproducibility demanded by modern diagnosis with automatic tube-current regulation are satisfied (see H. A. G. Hazeu and J. M. Lederboer, A universal Apparatus for X-ray Diagnosis, Philips techn. Rev. 6, 12, 1941.)

heating the filament had to be reduced to 0.5 sec. also, and this was achieved by making the filament longer and thinner, since the times, necessary to become hot, are proportional to the diameters of the filaments, given the same power.

Let us finally consider the employment of the tungsten anode for fluoroscopy. We have already pointed out, that it is no longer necessary to keep the basis temperature of the anode low. During fluoroscopy, there is no objection to the anode becoming red hot. The heat radiation of the anode then becomes so great that, without any further artificial means such as increasing the radiating surface, a very considerable power can be dissipated. In the case of the tube, shown in fig. 1, with a continuously applied power of 200 W the tungsten anode takes on a temperature of 800 °C, while tungsten can be made still much hotter without objection, especially since in the fabrication of the tube it was outgassed at a considerably higher temperature. The fluoroscopic power is therefore actually not determined by the anode temperature, but by the temperature taken on by the metal housing of the tube, which must finally give off the heat it receives to its surroundings. If it is desired, that the housing shall not become hotter than for instance 60 to 70 °C, so that it can be touched without burning the fingers, the fluoroscopic power for the very small housing according to fig. 2 is limited to about 150 W, as stated above.

We have already pointed out that, with a higher basis temperature, the margin to the highest permissible focus temperature becomes smaller. Therefore, if immediately after fluoroscopy one passes over to photography, as is the rule in stomach examinations, mentioned above, and is generally assumed as a basis for the determination of the loadability of every X-ray tube, the hot tungsten anode is at a disadvantage, compared with the cooler copper anode. This fact has been taken into account in making the tungsten-anode tube the equivalent of the old „Rotalix” tube by a corresponding extra increase of the radius of the anode. The loadability plotted in fig. 5 holds for a tungsten copper anode with the basis temperature 450 °C and a tungsten anode with the basis temperature 800 °C¹²).

It is of course unavailable that, when the tungsten disc is red hot, a certain flow of heat will also take place along the axle bearing the disc toward the rear and to the copper rotor (see fig. 1). Care must thus be taken, that the rotor does not become

too hot. To that end the heat transference by radiation from the turning rotor to the stationary anode support is increased by providing good radiating surfaces (blackened). An attempt might also be made to keep the flow of heat as small as possible, by making the axle long and thin and choosing as material a poor heat conductor. However, one is not absolutely free in this respect, as on the one hand the axle must be sufficiently stiff, *i.e.* sufficiently short and thick, to transmit the driving couple and not to fall into transverse vibrations, on the other hand also the axle (at least the front end of it) must be able to be outgassed at a high temperature. A not too high heat resistance of the axle has the advantage, that the heat capacity of the copper rotor can contribute to the improvement of the loadability of the anode in exposures with very long exposure times.

Prospects for further development

Summarising, we may say that the combination tungsten-copper offers indisputable advantages when it is desired to take full advantage of the gain involved in the principle of the rotating anode; that, however, the employment of tungsten alone makes it easier, as far as fabrication is concerned, to satisfy the requirements of general use in normal X-ray diagnosis. Only for those special purposes, where exposure times longer than a few seconds are necessary, namely for röntgencinematography, is the tungsten-copper anode so plainly superior, that it is clear, that in this field it will certainly maintain its usefulness, perhaps in an improved form. In the meantime the development of the tungsten anode is by no means at an end.

In spite of the fact that, as has already been explained, there can be no question of its competing with the tungsten-copper anode as far as heat capacity is concerned, it is quite possible to achieve a certain progress in that direction and thus to improve considerably the loadability of the tube for long exposure times. One path in that direction is the working out of methods of sintering larger tungsten discs. Recently indeed we have succeeded in making tungsten discs of more than 350 g instead of the abovementioned weight of 225 g. The loadability for long exposure times is increased by more than 50 percent in this way. A tube with such an anode is now at least the equal of tube provided with a tungsten - copper - anode of 1 kg, up to about 8 sec.

Finally, we should like to consider the possibility that a relatively slight increase of the specific focus loading, at very short exposure times, will

¹²) It is also possible, as is sometimes done, to lower the basis temperature somewhat by making the tungsten radiate more easily, for instance by making it black.

become desirable. Although we stated in the introduction, that at present there is no great need of this, further development might rapidly alter this fact. The possibilities of increasing the loadability at very short times were briefly indicated in the introduction: increase of the diameter and of the speed of revolution of the anode. As to the first, that trump has already been practically played in passing over from the tungsten-copper to the

tungsten anode, since we gave the tungsten disc a larger diameter to compensate for the poorer heat conduction and the smaller margin between basis and focus temperature. As to the second, the very much lighter tungsten disc is again very much in the advantage, since it will be much sooner possible, in that case, to increase appreciably the already high speed of revolution (about 2900 r.p.m.).

FREQUENCY MODULATION

by Th. J. WEIJERS.

621.396.619

After a short discussion of the ways in which a carrier wave of high frequency can be modulated, for instance with an audible frequency, so that the latter is made suitable for radio transmission, the method of frequency modulation is explained more fully. The principles of the necessary transmitter and receiver are described. Finally special attention is devoted to the so-called non-quasi stationary phenomena and their part in frequency modulation. It must be kept well in mind, that for the frequency in the different expressions it is not permissible to use simply the momentary frequency.

An audible sound can never be transmitted by means of radio by being converted directly into electromagnetic oscillations, because much higher frequencies are required for radio transmission. It is thus necessary to transmit a high frequency oscillation as carrier wave and to impress upon this wave, in some way or another, the low-frequency rhythm of the sound to be transmitted. This is generally done, by modulating the amplitude of the high frequency oscillation according to the low-frequency sound rhythm, and one then speaks of amplitude modulation. Even at the beginning of the development of practical radio-transmission, however, it was a well known fact, that, in principle, modulation could also be realized, by keeping the amplitude of the high-frequency carrier wave constant and varying the frequency according to the low-frequency rhythm to be transmitted. It was at first thought, that in such a system of so-called frequency modulation, there would be the advantage of being able to transmit signals with a smaller band width in the high-frequency region, so that more transmitters could be included in the same frequency region of carrier waves.

However, after Carson¹⁾ had shown in 1922 that this was incorrect, no more serious attempts were made in the following years to apply a system of frequency modulation for radio broadcasting.

New life was lent to attempts in this direction in 1936, when Armstrong²⁾ showed, that frequency modulation can offer the advantage, that the influence of disturbances on the reception can be more strongly reduced, than is the case with the customary amplitude modulation. Since then frequency modulation has indeed aroused a steadily increasing interest and the characteristics and practical possibilities of this system have been thoroughly investigated. As we intend to deal often

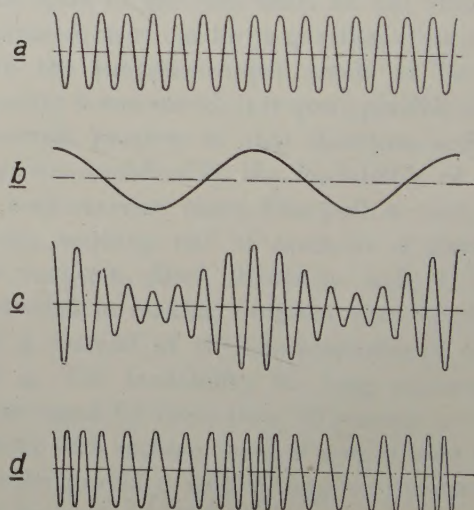
in the future with problems, which are connected with frequency modulation, the principle of frequency modulation will be further explained in this article, and then the characteristics will be discussed of the transmitters and receivers for such a system. A separate section will be devoted to the rôle of the non-quasi-stationary phenomena in frequency modulation. In a later article the two systems of amplitude and frequency modulation will be compared with respect to their degree of suitability for different applications.

The modulation of oscillations

A simple harmonic oscillation, such as is shown in *fig. 1a*, can be represented by:

$$V = A \cos (\omega_0 t + \varphi), \dots \dots (1)$$

where A , ω_0 and φ are independent of t . A is called the amplitude and $\omega_0 t + \varphi$ the phase of the oscillation. In order to make possible radio transmission of this oscillation, it is necessary, that the frequency $f_0 = \omega_0/2\pi$ should be much higher, than in the audible region. The transmission by radio of an oscil-



45286

Fig. 1. a) Unmodulated high-frequency oscillation.
b) Modulating low-frequency oscillation.
c) Amplitude-modulated oscillation.
d) Frequency-modulated oscillation.

¹⁾ J. R. Carson, Notes on the theory of modulation, Proc. I. R. E. 10, 57, 1922.

²⁾ E. H. Armstrong, A method of reducing disturbances in radio signaling by a system of frequency modulation, Proc. I. R. E. 24, 689, 1936.

lation with an audible frequency can thus never take place directly. A high-frequency oscillation is distorted in the rhythm of the low-frequency oscillation to be transmitted, in such a way, that radio transmission remains possible, i.e. that the resulting oscillation may be considered, as the sum of several sinusoidal components with sufficiently high frequencies lying in a relatively sufficiently narrow frequency region. In principle this can be done in two different ways: the amplitude A can be distorted and amplitude modulation is the result, or the phase $\omega_0 t + \varphi$ can be distorted and phase modulation results³⁾. As will be seen later, frequency modulation is a special case of this.

For the sake of simplicity, we shall assume for the time being, that it is desired to transmit a single sinusoidal low-frequency vibration $B \cos qt$, as drawn in fig. 1b. Thus the amplitude B and the phase qt of this vibration must be transmitted.

When this is done by means of amplitude modulation, the constant A in (1) is replaced by $A (1 + m \cos qt)$, m being proportional to B . In order to avoid distortion upon detection in the receiver, care is taken that m remains smaller than 1.

The vibration to be transmitted is represented by

$$V = A (1 + m \cos qt) \cos \omega_0 t. \quad (2)$$

The phase constant φ in (1) is of no importance here, and is therefore omitted in (2) for the sake of simplicity. This vibration is reproduced in fig. 1c. The varying amplitude of this curve is, in amplitude as well as in phase, an image of the vibration to be transmitted $B \cos qt$ (curve b). It must here be noted, that the zero points, i.e. the moments, when the vibration passes through zero, are not affected by the modulation.

Equation (2) can be resolved into three sinusoidal components:

$$A (1 + m \cos qt) \cos \omega_0 t = A \cos \omega_0 t + A \frac{m}{2} \cos (\omega_0 + q)t + A \frac{m}{2} \cos (\omega_0 - q)t.$$

$A \cos \omega_0 t$ is called the carrier wave, the two other terms the side-bands. The frequencies of these three components may satisfy the above conditions for radio transmission; the frequencies $\omega_0 - q$, ω_0 and $\omega_0 + q$ are high enough and lie in a frequency region which is relatively narrow enough when ω_0 is chosen sufficiently large. By relative width of the frequency region in this case is meant the quotient $2q/\omega_0$.

The three components in question can be represented in a vector diagram (fig. 2). The carrier

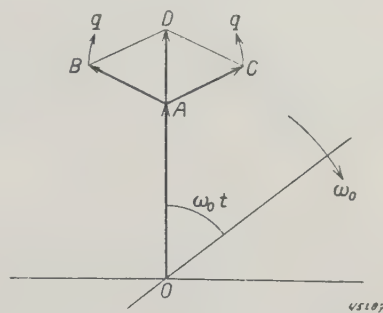


Fig. 2. Vector diagram of a sinusoidal amplitude-modulated signal OD . OA is the high-frequency carrier wave with the angular frequency ω_0 . AB and AC represent the side-band oscillation with the angular frequencies $\omega_0 - q$ and $\omega_0 + q$.

wave $A \cos \omega_0 t$ is represented by the stationary vector OA , the momentary value of which is equal to the projection of this vector on the line, which rotates around O with an angular frequency of ω_0 ; the two other components are represented by the vectors AB and AC , which rotate about the point A in opposite directions with an angular velocity of q . The amplitude modulated oscillation (2) is represented by the sum of these three vectors, i.e. by the stationary vector OD of varying length.

Such an oscillation can be generated in a transmitter, consisting of an oscillator with an amplifier. The oscillator voltage $A \cos \omega_0 t$ is applied to the grid of the amplifier valve. The amplification of this valve varies proportionally with the grid D C voltage or with the anode D C voltage. If one of these voltages is varied proportionally with $1 + m \cos qt$, the transmitter produces the required vibration proportional to $A (1 + m \cos qt) \cos \omega_0 t$.

One then speaks of grid or anode modulation, respectively, according to the method by which the amplitude-modulated signal is reproduced in the transmitter.

In the receiver, this oscillation is applied to a detector, for example a diode detector with leakage resistance. When an unmodulated oscillation, with a sufficiently large amplitude, is applied to this detector, a D C voltage occurs over the leakage resistance, which is approximately equal to the amplitude of the oscillation supplied. The frequency of the applied oscillation has practically no effect in this process of detection, provided it is sufficiently high. If the amplitude of the oscillation changes slowly, the voltage over the leakage resistance changes correspondingly. If the amplitude of the vibration is $A (1 + m \cos qt)$, the voltage given by the detector is $A (1 + m \cos qt)$ and the desired low-frequency oscillation $A m \cos qt$ is obtained.

³⁾ In principle, both amplitude and phase can be deformed, but little or no practical use is made of this possibility and we shall not discuss it here.

The D C voltage component A is of no interest, it is kept away from the following amplifier valve by a condenser.

If it is desired, to transmit the low-frequency oscillation $B \cos qt$ by means of phase modulation, the part φ of the phase in (1) is replaced by $m \sin qt$, so that (1) becomes ⁴⁾

$$V = A \cos (\omega_0 t + m \sin qt) \dots (3)$$

Like the amplitude-modulated oscillation, this oscillation also does not change sinusoidally with the time, and thus has no definite frequency. It is, however, possible to speak of a varying instantaneous frequency. To define this instantaneous frequency, we consider a non-sinusoidal oscillation:

$$y_1 = A \cos \{ f(t) \}, \dots (4)$$

where $f(t)$ changes arbitrarily with the time, and we look for a sinusoidal oscillation:

$$y_2 = B \cos (\omega t + \varphi), \dots (5)$$

which corresponds, as nearly as possible for the value of t considered, with the oscillation y , y_2 is then called the instantaneous frequency of y_1 . What is to be understood by „corresponds as nearly as possible“, is to a certain extent arbitrary.

Of the oscillation y_2 , one can choose the amplitude B , the frequency ω and the phase angle φ suitably; thus the oscillation can be made to satisfy these conditions. The most suitable choice is obtained, when one lets $A = B$, $y_1 = y_2$ and $\frac{dy_1}{dt} = \frac{dy_2}{dt}$ and then calculates B , ω and φ . One then finds:

$$\omega = \frac{df(t)}{dt} \dots (6)$$

and this frequency is defined ⁵⁾ as the instantaneous frequency ω_m of y_1 . When this definition is applied to the phase-modulated oscillation (3), one finds for the instantaneous frequency:

$$\omega_m = \omega_0 + mq \cos qt = \omega_0 + \Delta\omega \cos qt. \quad (7)$$

This varies therefore between $\omega_0 - \Delta\omega$ and $\omega_0 + \Delta\omega$. $\Delta\omega = mq$ is called the frequency sweep and $m = \Delta\omega/q$ the modulation index. Fig. 1d represents the phase-modulated oscillation (3). The amplitude does not change due to the modulation, but the zero points are shifted: the latter now lie sometimes closer together and sometimes farther apart than in the unmodulated oscillation.

The oscillation (5) is drawn in the form of a broken line curve in fig. 3 for the vicinity of one point. It is a sine curve (b) with the same amplitude

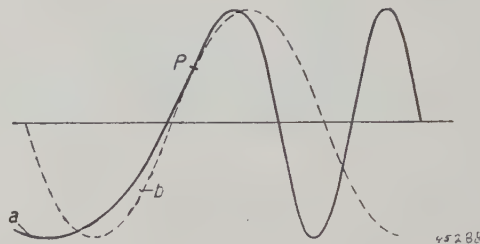


Fig. 3. Curve a represents a phase-modulated signal, while b is the sinusoidal oscillation whose frequency is equal to the instantaneous frequency ω_m of a at point P .

as the modulated oscillation (a), it passes through the point P in question and possesses the same tangent at this point.

Conversely, one may reason as follows. From (6) follows: $f(t) = \int \omega_m dt$. An oscillation with an instantaneous frequency $\omega_m(t)$ has the form $y = A \cos \int \omega_m dt$. If it is desired to transmit a low frequency oscillation $B \cos qt$, this can be done by an oscillation, whose instantaneous frequency is $\omega_0 + \Delta\omega \cos qt$, $\Delta\omega$ being proportional to B . This oscillation is thus $y = \cos \int (\omega_0 + \Delta\omega \cos qt) dt = \cos (\omega_0 t + \Delta\omega/q \sin qt)$.

This shows, why $\sin qt$ was taken in (3) and not $\cos qt$.

In the transmission of speech or music, it is not a single frequency, but a whole frequency region, which must be transmitted. The modulation system (transmitter and receiver) can now be so arranged, that the index m is independent of the frequency q of the modulating oscillation. Then the frequency sweep $\Delta\omega = qm$ is proportional to this frequency. Then one usually speaks of phase modulation in the narrower sense. It can, however, also be arranged so that not m , but the frequency sweep $\Delta\omega$ is proportional to the amplitude and independent of the frequency q of the modulating oscillation. The index $m = \Delta\omega/q$ is then inversely proportional to this frequency. One then speaks of frequency modulation. It has been found, theoretically as well as practically, that frequency modulation is preferable to phase modulation in the narrower sense. We shall therefore concern ourselves further only with frequency modulation.

Frequency modulation

A sinusoidal frequency-modulated oscillation can be resolved into sinusoidal components. In contrast to amplitude modulation, where this resolution gives one carrier wave and two side-band

⁴⁾ The reason why $\sin qt$ and not $\cos qt$ is used in (3) will appear later.

⁵⁾ The same definition is used by Helmholtz for acoustic vibrations.

frequencies, in frequency modulation an infinite series of oscillations is obtained, in which the frequencies of the successive components differ by an amount q and lie symmetrical with respect to the frequency of the unmodulated oscillation, which in this case also called the carrier wave.

Fig. 4 gives an idea of the amplitudes and fre-

region; and, moreover, that components occur, whose frequencies lie outside the region, which is covered by the instantaneous frequency. With a constant frequency sweep $\Delta\omega$ the region covered by the instantaneous frequency is more densely occupied by components, the smaller q is, i.e. the larger the modulation index m . This is not, however, surprising,

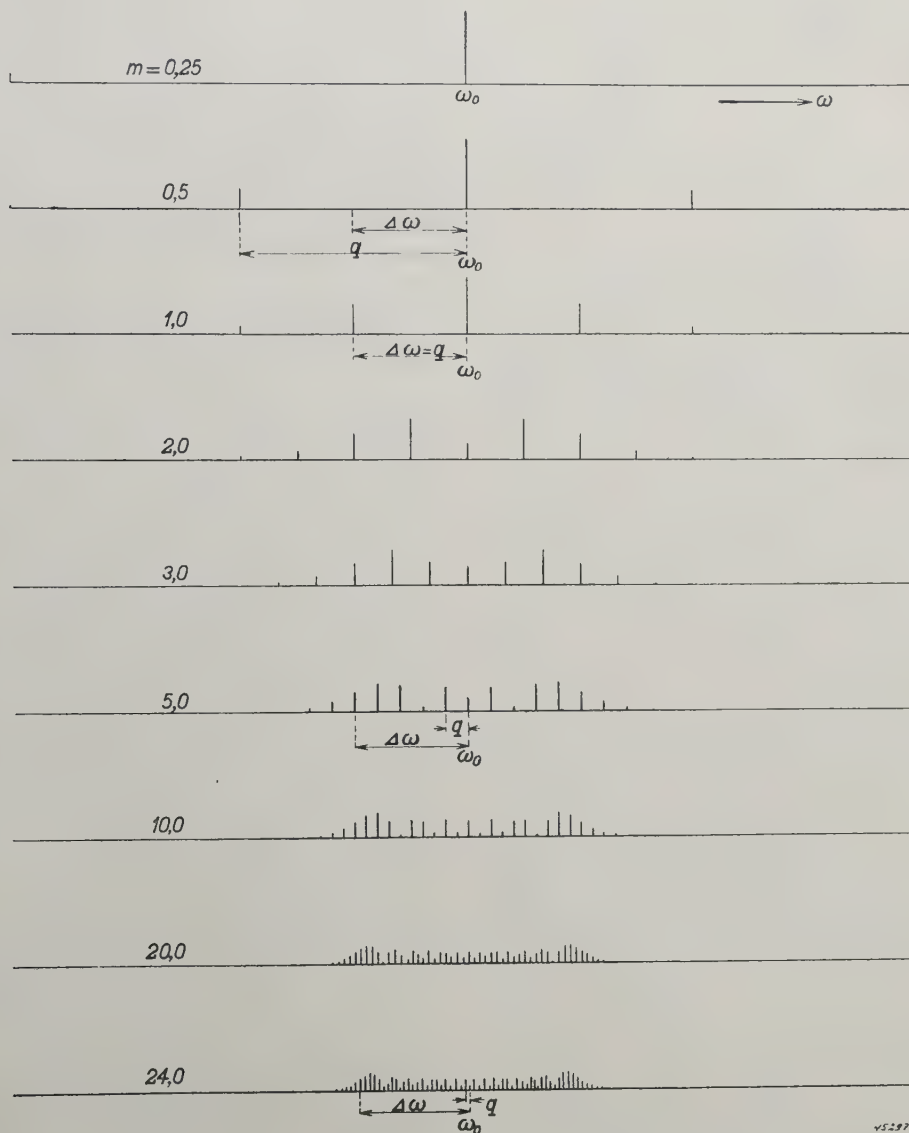


Fig. 4. Frequency spectra of frequency-modulated signals⁶⁾ with different modulation indices $m = \Delta\omega/2$ and constant frequency sweep $\Delta\omega$.

quencies of the components for different values of the modulation index m . It will be noted here, that although the instantaneous frequency ω passes continuously through the whole frequency region from $\omega_0 - \Delta\omega$ to $\omega_0 + \Delta\omega$, only a finite number of sinusoidal components occur in this frequency

when it is kept in mind, that the concept of instantaneous frequency is defined somewhat arbitrarily. Furthermore, it is clear from fig. 4 that, when $m > 1$, and thus $\Delta\omega > q$, the amplitudes of the components, whose frequencies differ more than about $3/2 \Delta\omega$ from the carrier frequency are very small and may, be neglected practically. If $m < 1$ and thus $\Delta\omega < q$, the first, at most the first two components on either

⁶⁾ Balh. van der Pol, Proc. J. R. S. 18, 1201, 1930.

side of ω_0 are of importance. In this case, however, no components at all, caused by the modulation, lie in the region covered by the instantaneous frequency. If m is small (for instance $m < 0.25$), the signal is practically the same, as the carrier wave with two side-band frequencies. These three components are represented in fig. 5 by vectors in an ana-

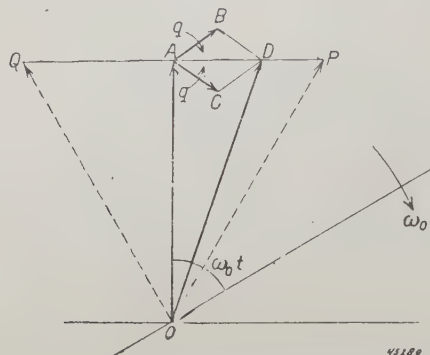


Fig. 5. Approximate vector diagram for a sinusoidal frequency-modulated signal whose modulation index m is a maximum of $1/4$. OA is the carrier wave with frequencies ω_0 ; AB and QC side-bands with frequencies $\omega_0 + q$ and $\omega_0 - q$; OD modulated signal.

logous manner to that in fig. 2 for amplitude modulation, but there is the difference, that the components of the side-band frequencies are rotated through an angle of $\pi/2$ with respect to the carrier wave. The sum of these three vectors is a vector OD , whose extremity moves along the straight line PQ , which is perpendicular to the vector OA of the carrier wave. The amplitude of the vector is thus not entirely constant. This is due to the disregarding of the other side-band frequencies. If on either side of the carrier wave an additional side-band frequency is included in the calculations, which can be represented by two smaller vectors, which rotate at double the angular velocity of the first two side-band frequencies, it may be seen from fig. 6

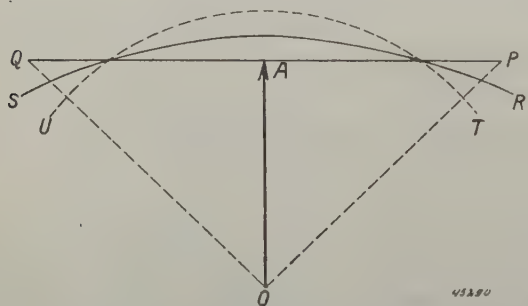


Fig. 6. The vector OA represents the carrier wave. On the straight line $PQ \perp OA$ lies the end point of the vector, which represents the sum of the carrier wave and one set of side-bands. On the curve RS lie the end points of the vectors, which represent the sum of the carrier wave and two sets of side-bands. On the arc of the circle TU lie the ends of the vectors, which represent the sum of the carrier wave and all (an infinite number) of side-bands.

that the sum of these five vectors gives a better approximation of the true vector, which, maintaining a constant length, fluctuates between the positions OT and OU , and which is obtained by including all the side-band frequencies (an infinite number) in the calculation.

In amplitude modulation a band width is necessary, which is equal to twice the highest audio frequency to be transmitted; in frequency modulation, for ideal transmission, strictly speaking one needs the whole infinite frequency region, practically, however, a frequency region equal to about three times the frequency sweep is necessary, if $\Delta\omega > q$, and equal to $2q$, if $\Delta\omega < q$.

Thus by the employment of frequency modulation no space in the frequency spectrum is saved, as was previously thought, but on the contrary more space is needed!

Transmitter for frequency modulation

A transmitter for frequency modulation can be arranged in different ways. Armstrong set about it as follows. An oscillator controlled by a crystal furnishes the carrier wave. This carrier wave is amplitude-modulated, by the low-frequency oscillation to be transmitted, in a push-pull modulator, so arranged, that it furnishes only the two side bands without the carrier wave. The carrier wave is then added again with a phase rotation of $\pi/2$ and a large amplitude, so that an oscillation like that of fig. 5 is formed. Care is taken that the angle $POA < \pi/12$, so that the amplitude variation remains small enough to be disregarded. The argument of $\cos(\omega_0 t + m \sin qt)$ represented by the varying vector OD , changes between $\omega_0 t + m$ and $\omega_0 t - m$. From this follows, that $\angle POA = m$. The condition that $\angle POA < \pi/12$ therefore expresses, that $m = \Delta\omega/q < \pi/12$ or $\Delta\omega < \pi/12 \cdot q$. For a low modulation frequency, for instance $q = 2\pi \cdot 30 \text{ sec}^{-1}$, $\Delta\omega < \pi/12 \cdot 2\pi \cdot 30 \approx 50$. For higher modulation frequencies $\Delta\omega$ is proportionally, but even for the highest modulation frequencies, which occur in praxis e.g. $q = 2\pi \cdot 15,000$, corresponding with the highest audible frequency, the frequency sweep is not greater than $2\pi \cdot 25,000 \text{ sec}^{-1}$.

By the method just described, phase modulation in the narrower sense would be obtained, because the modulation index m was assumed to be constant. In order, however, to obtain frequency modulation (i.e. the frequency sweep $\Delta\omega$ independent of the modulating frequency), before the amplitude modulation, the low-frequency signal is passed through a network, in which the amplitudes of the different frequencies are amplified proportionally

with the modulating frequencies. In order to obtain a large frequency sweep, which is desirable as will be discussed in a later article, the frequency of the modulated signal is multiplied in several successive stages, before it is transmitted. The carrier frequency ω_0 as well as the frequency sweep $\Delta\omega$ are thereby multiplied. The advantage of this method is, that a crystal-controlled oscillator is used, so that the carrier frequency is very constant.

In another method of modulation much used in transmitters, the capacity of the oscillation circuit is varied. If a condensor microphone is connected in parallel with the tuned circuit of an oscillator, the capacity of the circuit varies upon speaking into the microphone. The instantaneous frequency of the oscillator thereby also varies. If the variation of the capacity is small, compared with the total capacity of the circuit, the variation of the instantaneous frequency is approximately proportional to the capacity variation. Practically, however, there is an objection to including the microphone directly in the oscillator circuit. In order to avoid this, a so-called reactance valve is preferably connected in parallel with the oscillator circuit, which valve serves as variable capacity or variable self-induction⁷⁾. The action of the reactance valve is as follows. Between anode and cathode an A C voltage is applied with the same frequency as, but shifted in phase with respect to, the anode voltage. The grid voltage can be derived in a simple way from the anode voltage, for instance by introducing a resistance R between anode and grid and a capacity C between grid and cathode (fig. 7). Here

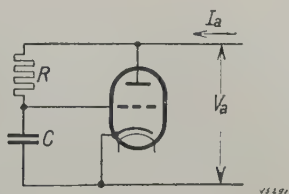


Fig. 7. Connections for the reactance valve. With the help of the resistance R and the condensor C the grid voltage is derived from the anode voltage V_a .

$V_g = V_a / (1 + jR\omega C)$. The anode A C is approximately $I_a = SV_g = V_a S / (1 + jR\omega C)$, if S represents the slope of the characteristic. The impedance measured between anode and cathode is now $V_a / I_a = 1/S + jR\omega C/S$, and therefore equivalent to a connection in series of a resistance $1/S$ and a self-induction RC/S . It is here assumed, that the impedance of the series connection of R and C is so high, that the current through this impedance may be

neglected compared with the anode A C. By varying the negative grid voltage of the valve the slope S is varied and thereby also the effective self-induction RC/S and thus the frequency of the oscillator in parallel with whose circuit the reactance valve is connected (fig. 8). The positions of R and C

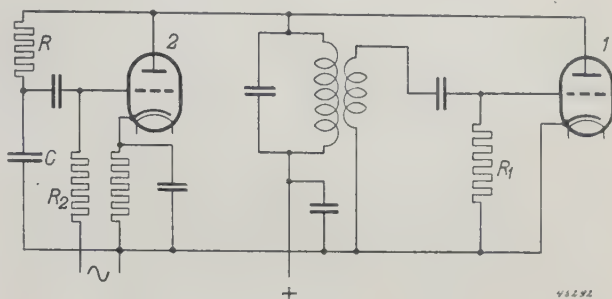


Fig. 8. Oscillator circuit with oscillator valve 1 and reactance valve 2/ R and C are introduced according to fig. 7. The modulating signal is fed to the grid over the large resistance R_f .

can also be interchanged, where upon the reactance valves becomes equivalent to a resistance and a capacity in series. It is also possible to use a self-induction L , instead of the condenser C .

When the modulating low-frequency voltage from the microphone is now applied to the grid over the resistance R_f , the desired frequency-modulated signal is obtained in the oscillator circuit. The frequency of the frequency variation of this is equal to the frequency of the modulating signal. With suitable characteristics of the valve employed and suitable dimensions of the other elements of the connections, the frequency sweep is sufficiently nearly proportional to the amplitude of the modulating signal, provided it does not become too large.

Receiver for frequency modulation

In the receiver the frequency modulated signal must again be converted into a low-frequency oscillation, which is a faithful copy of the voltage at the microphone. The detector (diode), employed in the amplitude-modulation receiver, reacts only to the amplitude and not to the frequency of the high-frequency signal applied and cannot therefore be used directly in frequency modulation, where the amplitude is constant. Therefore in the frequency modulation receiver the frequency-modulated signal is first converted into an amplitude-modulated signal and then applied to a diode detector. This conversion of a frequency-modulated signal into an amplitude modulated signal is accomplished by means of a frequency detector i.e. a network, which furnishes a high-frequency voltage, whose amplitude is sufficiently nearly a linear function of the frequency of the applied voltage,

⁷⁾ Compare also: Philips techn. Rev. 4, 85, 1939.

if the amplitude of this applied voltage is constant. For this purpose may be used, for example, connections which consist of a resistance, a self-induction and a capacity in series. The resonance curve, which represents the voltage E over these connections as a function of the frequency, with constant current through the circuit, possesses a section, which is approximately straight (AB in fig. 9). If the frequency of the current varies over the region corresponding to this almost straight part, the variation of the voltage E over the circuit is proportional to this frequency variation. The sine curve a represents the instantaneous frequency ω_m of the modulated signal as a function of the time t ; the sine curve b represents the instantaneous amplitude E of the high-frequency voltage over the circuit as a function of the time t . Except for a constant, this amplitude is proportional to the instantaneous frequency of the modulated signal minus the carrier frequency. This voltage, which thus contains both frequency modulation and amplitude modulation, is applied to a diode detector, which does not react to the frequency modulation, but to the amplitude modulation of this voltage. In this way, from a frequency modulated signal $\cos \{f(t)\}$ with a carrier frequency ω_0 , one obtains a low-frequency signal, which is proportional to the instantaneous frequency minus the carrier frequency i.e. to $\{df(t)/dt\} - \omega_0$.

If we prolongue the nearly straight portion AB of the characteristic in fig. 9 until the frequency axis, this axis is cut in a point, lying at a distance q_1 , of the central frequency ω_0 . Moreover, we see in fig. 9, that the voltage E possesses an amplitude modulation with a modulation depth $\Delta\omega/q_1$. The ideal would be, that the characteristic fol-

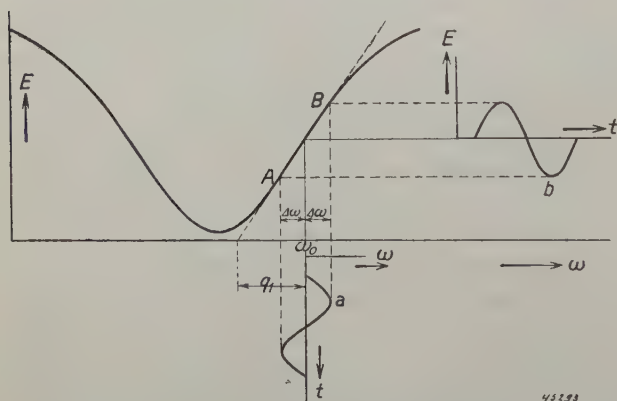


Fig. 9. High-frequency amplitude of the voltage E over a connection in series of a resistance, self-induction and capacity, as a function of the frequency ω . If we assume, that the instantaneous frequency varies sinusoidally according to a with the time t , we obtain the relation given by b between the instantaneous amplitude E of the modulated high-frequency signal and the time t .

lowed indeed the whole of the straight dotted line; in this case the signal, arising after frequency-detection, would have obtained an amplitude modulation of 100% as a consequence of the frequency-sweep q_1 , of the original frequency-modulation.

Another often used circuit for frequency-modulated signals is shown in fig. 10. In the anode circuit

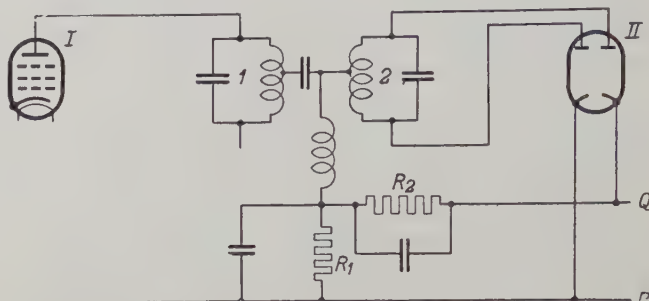


Fig. 10. Detector circuit for frequency-modulated signals. I last intermediate-frequency valve; II duodiode, 1 and 2 are the primary and secondary circuits of the frequency detector. Between the points P and Q we obtain the detected low-frequency signal.

of the last intermediate frequency, valve I of the receiver a network is included, which consists of two inductively coupled circuits, each tuned to the frequency of the carrier wave of the frequency-modulated signal to be received. For this frequency, the voltages over the primary circuit 1 and the secondary circuit 2 differ in phase by $\pi/2$; for higher frequencies this phase shift changes in one direction, for lower frequencies in the other. The middle of the secondary circuit 2 is connected with a point of the primary circuit 1. To each of the two diodes, of which valve II consists, a voltage is applied consisting of the sum of a part of the voltage of the primary circuit 1 and half of the voltage of the secondary circuit 2. These voltages are represented in fig. 11 in a vector diagram. If the signal is unmodulated, OA is the primary voltage, AB and AC are the two halves of the secondary voltage.

The voltages on the two diodes are thus OB and OC . The D C voltages over the resistances R_1 , and R_2 of fig. 10 are equal, since the vectors OB and OC are equal in length. Between the points P and Q of fig. 10 one obtains the difference between these two voltages, which is zero in this case. If, however, the frequency, of the signal applied, changes, the vectors AB and AC rotate, for instance to AB^1 and AC^1 . The voltages over R_1 and R_2 are no longer equal, but $OB^1 > OC^1$. The voltage between points P and Q is now equal to the difference between the lengths of the vectors OB^1 and OC^1 . If the frequency of the signal applied varies in the opposite direction the sign of the voltage between P and Q changes.

If the applied signal is frequency-modulated, there therefore occurs between P and Q a low-

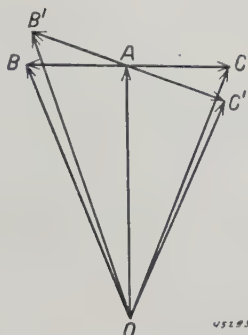


Fig. 11. Vector diagram for the frequency detector according to fig. 10. OA is the voltage, on the primary circuit 1. OB and OC are the voltages, which the carrier wave furnishes to the two diodes, of which valve II consists, OB' and OC' are furnished at a frequency, which deviates slightly from the carrier frequency ω_0 .

frequency A.C. voltage, which is an image of the modulation. By suitable choice of the coupling between the two circuits, of the ratio of the primary to the secondary voltage and of the damping of the circuits, the voltage between P and Q can, in a sufficiently wide frequency region, be made very nearly proportional to the frequency deviation from the carrier frequency. The desired low-frequency signal is therefore obtained free of distortion. Fig. 12 gives for such a frequency detector the voltage between P and Q as a function of the angular frequency ω of the signal applied. This curve approximates to a straight line, closely enough between the frequencies ω_1 and ω_2 , so that the

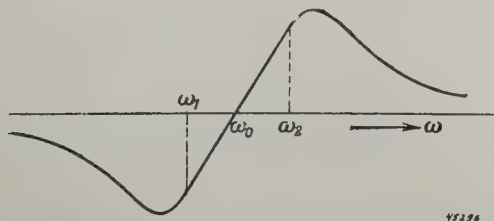


Fig. 12. The detected voltage between the points P and Q of fig. 10 as a function of the frequency ω . In the region around the carrier frequency ω_0 the curve is practically a straight line between ω_1 and ω_2 .

detector can be used satisfactorily for a carrier frequency ω_0 and a frequency sweep $\Delta\omega$ which must be slightly smaller than $\omega_0 - \omega_1 = \omega_2 - \omega_0$, as will be discussed in the following section.

Non-quasi-stationary phenomena

In the foregoing, it has always been tacitly assumed, that the voltages and currents in networks behave, under the influence of signals with varying

frequency, in the same way as at constant frequency, thus that the instantaneous frequency may simply be substituted for the frequency in the expression for the impedance. The fact, that this is at least doubtful, follows immediately from the somewhat arbitrary definition of instantaneous frequency. Carson and Fry⁸⁾ have shown, how it is possible, to calculate the behaviour of arbitrary networks under the influence of a voltage with variable frequency. This exact calculation is quite complicated, but simple practical conclusions can be drawn from it.

When an A.C. voltage $V = V_{max} \cos \omega t$ is connected at the moment $t = 0$ with a resistance r and a self-induction L in series, a current

$$i = i_{max} \cos(\omega t - \varphi) - i_{max} \cos \varphi \cdot e^{-\frac{r}{L}t}$$

flows. This current can be divided into two parts: a "quasi-stationary" part $i_{max} \cos(\omega t - \varphi)$, which continues to exist as long as the external voltage acts, and a transient current

$$-i_{max} \cdot e^{-\frac{r}{L}t} \cos \varphi,$$

which practically disappears after a certain, usually very short time. In the case of a more complicated network also the current always consists of a stationary part and a transient current, which decreases to zero. Also when the amplitude or the frequency of the voltage applied changes, a temporary component of the current appears. With arbitrarily changing voltages therefore, one is concerned with such temporary components. If the change in frequency takes place so slowly, that the temporary component has already become very small before any appreciable frequency change has again taken place, *i.e.* in such a way, that the network has time to adjust itself to the new situation, these exponentially decreasing components are scarcely noticeable, and the currents and voltages in the network can be measured practically with sufficient accuracy, as if one were dealing with a constant frequency, namely as if the frequency at every moment were equal to the instantaneous frequency. In this case the network behaves as if quasi-stationary. In any case this method, of considering it quasi-stationary, gives a first approximation of the phenomena occurring. The extent, to which this approximation produces sufficiently accurate results in a practical case, depends upon the modulating frequency q , the frequency sweep $\Delta\omega$, and thus also on the modulation index m , as well as on the dam-

⁸⁾ J. R. Carson and T. C. Fry, Variable frequency electric current theory, B. S. T. J. 6, 513, 1937.

ping of the network. When in this connection the spectra of fig. 4 are again considered, it is seen that although the instantaneous frequency covers a region $2\Delta\omega$ continuously, the oscillation is composed of components with discrete frequencies. With a large values of m , i.e. with a slow variation of the frequency, the region of the instantaneous frequency is indeed more densely occupied, than at small values of m , and thus more closely approaches a continuous spectrum, as would be expected according to quasi-stationary considerations. Moreover, with large values of m a relatively smaller region lies outside the region of the instantaneous frequencies than with smaller values of m . If $m < 1$, except for the carrier wave, not even a single component lies in the region covered by the instantaneous frequency. From this it may be seen again, that the quasi-stationary method of consideration leads to more accurate results the larger m is, and thus with a given value of $\Delta\omega$, the smaller q is.

In the amplifier stages of transmitter and receiver tuned circuits are employed, as a rule band filters, consisting of two coupled circuits. From the quasi-stationary method of consideration, it would follow, that the band width of these filters is sufficient, when a frequency region $2\Delta\omega$ is uniformly amplified. From the spectra of fig. 4, however, it follows that this band width is insufficient. If $m < 1$ a band width of at least $2q$ is necessary, otherwise the whole

modulation might even disappear. If $m > 1$ a band width of about $3\Delta\omega$ is necessary; otherwise too many side-band frequencies are cut off, which would cause distortion in the modulation and therefore also in the low frequency signal detected.

From this it is evident, that the quasi-stationary method of consideration, which led to the concept of instantaneous frequency, can easily lead to incorrect conclusions in the case of frequency modulation. The incorrect concept, mentioned in the introduction, that a frequency-modulated signal with a small frequency sweep would require only the region in the frequency spectrum, which is covered by the instantaneous frequency, was a result of the quasi-stationary method of consideration, which does indeed give a useable approximation, when the modulation index m is large, but which leads to absurd results when m is small.

After the above discussion of the way, in which frequency modulation can be realized for the transmission of low-frequency signals on a high-frequency carrier wave, there remains the problem, as to the cases in which such a system can be successfully used for radio transmission. To what degree frequency modulation offers advantages over the usual broadcasting system of amplitude modulation will, however, be discussed in a later article.

THE FORMATION OF STEREOPHONIC IMAGES

by K. de BOER.

534.621.3

It is possible to construct an installation for sound amplification or reproduction, such as is used for orchestra music, sound film or stage plays, in such a way that the „acoustic image” heard by the audience always coincides with the visual image which is seen. In particular the conditions are here discussed under which the sound must be recorded in order to attain the desired result. Several more particulars are also given about the technique of reproduction which has been dealt with previously, while it is also pointed out that stereophonic reproduction, without the source of the sound being visible, already constitutes a considerable improvement.

Equipment for sound amplification is installed in a hall in which a small orchestra performs. In the usual way a microphone is hung somewhere in front of the orchestra and the sound received by the microphone is amplified and reproduced by a loudspeaker. When the attention of a member of the audience is especially attracted first by the violon at the left and then by the piano more to the right, he will often feel troubled, that the music which he hears, does not come from the direction in which he sees the violin or piano, but always from the same direction, that of the loudspeaker. Sound reproduction at the cinema is usually subject to this objection. Here also, the sound always comes from a single direction, even though the source of sound (the speaker) moves about on the screen. This often produces an unnatural effect.

It is also a remarkable fact that in those cases where sound reproduction is employed, without the audience being able to see the original sources of the sound (or pictures of them), a considerable improvement of quality can be attained by applying stereophonic reproduction. Even the reproduction by means of two „ordinary” loudspeakers connected in series, placed at some distance from each other, already provides an improvement, which is sometimes made use of in the case of the radio. For the present, we shall confine ourselves to mentioning this fact, and in the following devote our special attention to those cases, where the audience can see the source of sound or a picture of it and where it will be expected, that the sound reproduced will come from the direction, in which the source of that sound is seen.

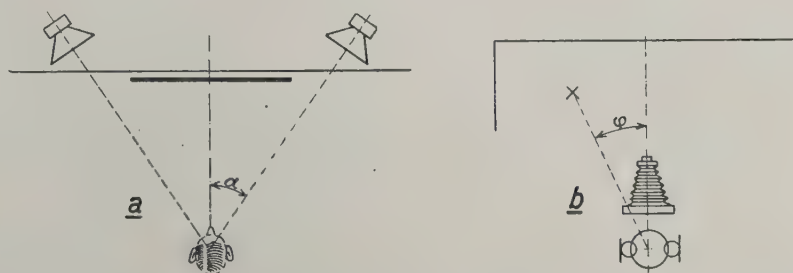


Fig. 1. *a*) Arrangement for stereophonic reproduction in a cinema. Two loudspeakers are placed one on either side of the projection screen and are observed within an angle $2a$ by the listener.

b) Arrangement of the artificial head in the studio. The perpendicular bisecting planes of camera and artificial head are coincident. The speaker stands for instance in the direction φ .

With the stereophonic method of sound reproduction these objections can be removed and the sound reproduced is always heard from the same direction, as that in which its source is seen. It is important, that even on those persons, who are scarcely aware of the objections mentioned, stereophonic reproduction, nevertheless, often produces a more pleasing impression than ordinary reproduction.

Principle of the arrangement for stereophonic image forming

The principle by which the coincidence of sound image and visual image can be attained has already been sketched in a previous article¹⁾ for the case of the cinema, particular attention being devoted

¹⁾ See the article: Stereophonic sound reproduction, Philips techn. Rev. 5, 107, 1940.

to the technique of reproduction. We shall review it briefly here, see *fig. 1*. For sound recording an "artificial head" (a sphere) is set up in the studio, which has two microphones at the position of the ears. The sound recorded on the film for each of the two microphones separately is reproduced in the well-known manner *via* separate channels, in which an amplifier is included, by a loudspeaker. The two loudspeakers are placed at either side of the screen in the projection room. It must be pointed out particularly, that this method of reproduction is simpler and pleasanter than that in which each member of the audience must wear headphones, while by calculation and experimentally it was found, that the accuracy of the sound image need not suffer appreciably. In this article therefore, we shall not devote our attention so much to reproduction, but to the requirements made of the method of sound recording. The question thus is, how must the sound be recorded in the case of the orchestra and in analogous cases (stage, film), in order to obtain an acoustic image, which is true to nature when the above-mentioned method of reproduction is employed. It is found, that for this purpose both the size and the position of the artificial head must be chosen according to a definite rule. Before we state this rule and discuss its application, we shall outline the experiment from which it results.

In a recording room, see *fig. 2*, an artificial head and a moveable source of sound *G*, are set up. In the recording room *R*, the observer *W* observes the sound image of *G* between the two loudspeakers *L*₁ and *L*₂ at a distance *u* from the middle.

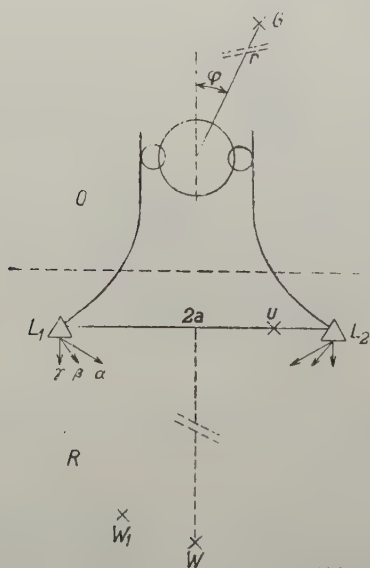


Fig. 2. In the recording room *O* an artificial head and a moveable source of sound *G*, are set up. In the recording room *R*, the observer *W* observes the sound image of *G* between the two loudspeakers *L*₁ and *L*₂ at a distance *u* from the middle.

The distance $2a$ between the loudspeakers, the base, can be chosen differently, for instance normal stage width of 4.5 m, width of the projection screen, etc.

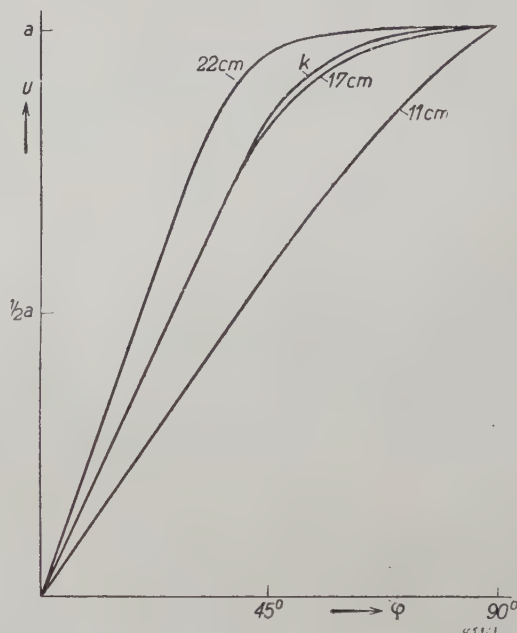


Fig. 3. Upon recording the sound from a source *G* (*fig. 2*) placed at an angle φ with the plane of symmetry of the artificial head *u* represents the deviation from the middle of the sound image as observed by *W* in the projection room. The various curves are for the different sizes of artificial head as indicated in the figure. *k* stands for a perfectly shaped head.

About in the middle of the hall, in our case at a distance of 9 m, the observer *W* is seated. In the recording room a source of sound is placed and its position fixed by the distance *r* and the angle φ . In the projection room *W* then hears the sound image between the loudspeakers at a distance *u* from the middle. We now determine this deviation *u* at different values of φ and *r*.

It is found, that the deviation *u* of the sound image depends only little upon *r*, so that we obtain in fact only a relation between φ and *u*, thus between the angle, measured from the plane of symmetry, at which the artificial head „hears” the source of sound and the deviation of the sound image from the middle.

This relation, which differs for different sizes of artificial head, is given in *fig. 3* for three different cases:

1. For an artificial head (sphere) 22 cm in diameter, which is slightly larger than a normal head;
2. For a well modelled head, and for a sphere of 17 cm diameter, which is about the same size as the head;
3. For a sphere 11 cm in diameter.

The graph shows that, for the 22 cm sphere, the relation between φ and u is linear as long as φ is not greater than 45° . This linear relation, in which equal angular rotations φ correspond to equal displacements u , gives a sound image, whose accuracy and naturalness are more than sufficient for practical purposes. As soon, however, as the source of sound has a greater angular deviation than 45° from the centre, in the case of the 22 cm sphere, the sound image always lies close to one of the loudspeakers, so that we obtain a much too compressed and distorted sound image which does not give a pleasing impression. (For example all the musicians of a large orchestra who are observed by this artificial head in directions deviating by more than 45° from the centre would in the acoustic image appear to be sitting in a lump, while the others would be at normal distances apart).

It is found further that, with the small sphere, we can reproduce a larger range of angles, namely to approximately 90° from the centre, undistorted on the base. This can easily be understood, because with a given angle φ , the ratios of intensities and differences in time are smaller with the small sphere than with the large, as that the corresponding deviations u are smaller and a wider range of angles can be accommodated on the given base²⁾.

For the relation between the size of the artificial head and the width of the range of angles, which can be reproduced undistorted on the base, we can further deduce from fig. 3 the following rule of thumb: (diameter of artificial head in cm) $\times$ (angle to be projected in degrees) = 2000.

This rule of thumb is very useful in solving problems like the following. Stereophonic sound amplifying apparatus is installed in a hall in which an

must be chosen for the artificial head and where should it be placed?

The correct position is simply found. Is is the same as the position of the single microphone in ordinary sound amplification, and certain rules for this are known from practice. When the artificial head, which is placed according to these rules „sees” the orchestra within an angle of 100° , for correct image forming we must, according to the rule of thumb, chose a diameter of 20 cm, see fig. 4a. We then obtain an image like that shown in fig. 4b. Our problem is thus solved sufficiently accurately for practical purposes; the objection described at the beginning of this article has been met. Analogous considerations hold for the formation of the acoustic image of a large orchestra, an open-air play and such cases. In the meantime stereophonic reproduction has been employed in practice only for the cinema on a large scale, while for the other cases it is still very much in its initial stages.

In connection with these figures we shall discuss several particulars in more detail.

Since fig. 3 is valid for the observer W , it might be thought that, by an observer at any other point in the hall, the sound image would also be observed at another point u . It would lead us too far afield to deal with this question in detail and we shall confine ourselves to a single aspect. When W moves along the line of symmetry the sound image scarcely moves. The displacement may, however, become somewhat greater, when W moves to one side to a point W_1 (see fig. 2). At this point the influence of L_1 , is dominant and the image shifts somewhat to the left. This could, however, be avoided by giving L_1 and L_2 , a certain directional effect, i.e. by providing that the same sound energy is not radiated

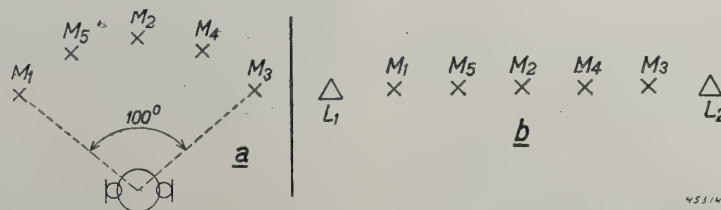


Fig. 4. a) The artificial head „sees” the five musicians of an orchestra within an angle of 100° . Upon correct choice of the diameter of the artificial head an acoustic image is formed as sketched in fig. 4 b) whose accuracy is sufficient in practice. The loudspeakers L_1 , and L_2 are actually placed on either side of the orchestra.

orchestra consisting of five musicians is performing. A loudspeaker for the sound reproduction is mounted on either side of the orchestra. How must the sound now be recorded, i.e. which diameter

in the directions α , β , γ (fig. 2), but that, as the length of the arrows indicates, there is a decrease in the intensity radiated in the order given. In this way the dominating influence of L_1 (and the diminished influence of L_2) can be compensated, and an undistorted image can be obtained throughout practically the whole hall.

²⁾ See the article: Stereophonic sound reproduction, Philips techn. Rev. 5, 107, 1940.

Width regulator

It is also evident, from the experiment described above, that if the base $L_1 L_2$ had been chosen longer than the width of the orchestra in question, we would still have obtained an undistorted, but at the same time an enlarged acoustic image.

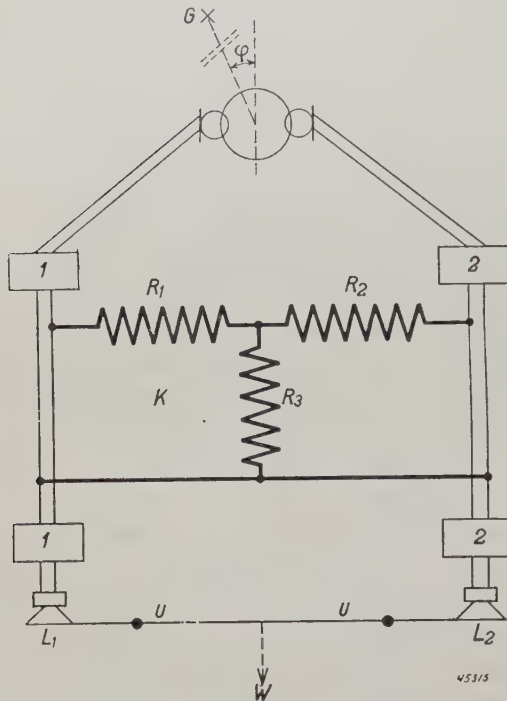


Fig. 5. Diagram of an arrangement for stereophonic reproduction with adjustable width of sound image. The sounds received by two microphones on either side of an artificial head in the recording room are reproduced by the loudspeakers, L_1 and L_2 . The observer W then observes an "acoustic image" of the recording room between L_1 and L_2 . By means of the coupling element K , indicated by heavy lines, the width of this image can, if necessary, be changed. In this way it is always possible to attain satisfactory correspondence with a visual image also observed by W .

Thus, when it is desired to „project” images of objects of different sizes (large orchestras, stage plays or films, small orchestras) in the same hall, the loudspeakers should in each case be placed at a different distance apart. This difficulty can now be entirely eliminated by a slight modification in the electrical connections, which are indicated by heavy lines in fig. 5. We shall show that, even though the loudspeakers are located permanently at the greatest distance apart which will ever be necessary, with the connections of fig. 5 the apparent distance between the loudspeakers can be decreased in a simple way, so that the sound image can always be adjusted to the natural size.

The coupling element K indicated by heavy lines consists of three variable resistances, namely R_1 and R_2 , which are always taken equal, and R_3 .

(At the same time it is also indicated in fig. 5 that for practical reasons, which we shall not go into, the amplifier in each channel is constructed in two parts, a preamplifier and an end amplifier.) Due to the presence of K part of the current from preamplifier 1 enters the „wrong” channel (namely 2), while at the same time the same fraction of the current from preamplifier 2 also enters the wrong channel (namely 1). Suppose now that the sound radiated by a source G is received by the left-hand microphone at the moment t_1 , and a moment later at t_2 by the right-hand microphone. Part of the sound energy taken up at moment t_1 , by the left hand channel, after amplification to a certain amount I_1 is radiated by the left hand loudspeaker also at the moment t_1 . (The transit time of the sound converted into electrical vibrations can be neglected). Another part, however, is fed, *via* the coupling element, to the second end amplifier, whereupon the right hand loudspeaker radiates an energy pI_1 , also at the moment t_1 , the fraction p being regulated by the varying of R_1 ($= R_2$) and R_3 .

In fig. 6 this situation is sketched. A time axis is indicated in the vertical direction. The arrows pointing to the left are proportional in length to the energies radiated by the left hand loudspeaker, those pointing to the right to the energies radiated by the right hand loudspeaker. At the moment t_1 , therefore we have I_1 , on the left and pI_1 , on the right. For the sound, which is received by the right hand channel at t_2 , the case is analogous. We then find that at t_2 I_2 is radiated by the right hand loudspeaker and pI_2 by the left hand one, p being the same fraction as before, because of the complete symmetry of the connections. This is also indicated in fig. 6; at t_2 we have thus I_2 to the right and pI_2 to the left. We now determine the position of the sound image observed by W . The ratio I_1/I_2 is the

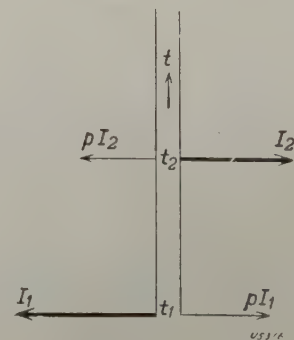


Fig. 6. The arrows toward the left indicate the intensity of the sound emitted by the left-hand loudspeaker at the moments t_1 and t_2 . The arrows toward the right are for the right-hand loudspeaker. The arrows drawn with a thin line represent the effect of the coupling K of fig. 5.

same as would be radiated by L_1 and L_2 in the absence of the coupling K , since I_1 and I_2 have both lost the same fraction. The time difference $t_1 - t_2$ is also unchanged by K . The combination $(I_1/I_2, t_1 - t_2)$ alone, would therefore be observed by W as if it came from a sound image which shows a deviation u to the left, corresponding to the angle φ according to fig. 3. However, W also observes the intensities pI_1 and pI_2 which are in the same ratio I_1/I_2 and have the same time difference $t_1 - t_2$, but, in contrast to the above, the highest intensity (pT), comes from the right hand loudspeaker. The combination $(pI_1/pI_2, t_1 - t_2)$ is thus observed by W as an image, also with a deviation u , but now to the right, and weaker in intensity by a factor p . W thus finally observes the sound image which would occur if two loudspeakers at a distance u to the left and right of the middle radiated intensities with the ratio p , of which the left-hand one is the stronger. According to the measurements already published³⁾ the sound image deviates from the middle toward the left by a fraction of u determined only by p . We thus reach the conclusion, that due to the „cross talk” caused by K the deviation from the middle of every sound image is decreased in the same ratio (determined by p), so that we observe a reduced but undistorted image.

Fig. 7 gives some examples, how the image of an artificial head can be affected by a width regulator. The deviation of the sound image from the centre has been recorded as a function of the change of φ , with regard to the plane of symmetry of the artificial head. The dotted line *I* is the image, measured without a width regulator, for an artificial head and was drawn as smooth as possible through a great number of observations.

From the experimental curve *I* we have, for two distinct positions of the width regulator, calculated the drawn curves *II* and *III* according to the here explained theory. The measurements for these cases coincide satisfactorily with the drawn curves.

The regulation of this reduction is carried out in practice with a single knob which in a number of positions switches in suitable values of R_1 ($= R_2$) and R_3 , while these values are, moreover, so chosen that the total intensity of the sound does not change upon regulation (width regulator).

Left-right regulator

It is clear that the two reproduction channels *I* and *2* of fig. 5 must not cause any alteration in the

intensity relations and the time differences of the sounds received. As for the time differences, there is no danger of this happening with the connections employed; special measures must, however, be taken against a change in the correct intensity ratios. The two fixed amplifiers and the two loudspeakers can be made mutually identical once for all. A difference in the sensitivity of the recording microphones due to which a displacement of the whole sound image toward the left or right would occur upon reproduction, is usually combatted in a different way. Since different microphones are often used with different recordings, because for example of the different sizes of the artificial head, it is in practice simplest to include a potentiometer in each channel, and to increase the energy taken up in one channel somewhat, if necessary, and at the same time to decrease that in the other channel. This has the same effect as a corresponding alteration of the sensitivity of the microphones. The sound image can then be put back in the right position without a change in the total energy (*left-right regulator*).

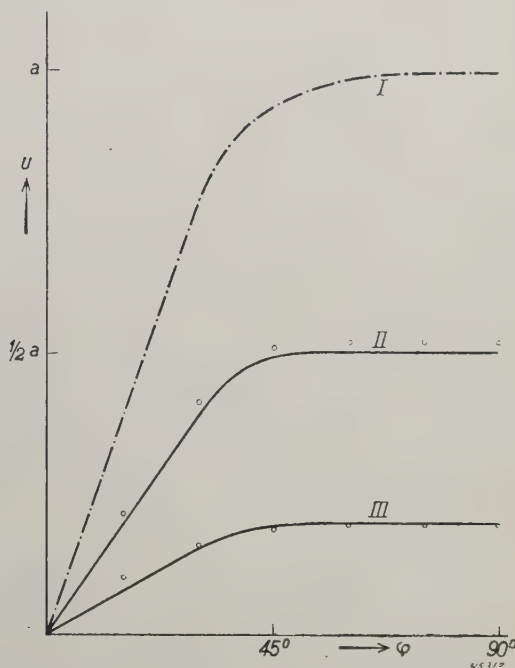


Fig. 7. Upon recording the sound from a source G (fig. 2), which is placed at an angle φ to the plane of symmetry of two separately mounted microphones, u gives the deviations of the sound image from the middle, as it is observed by W in the projection room. The various curves are for the different distances between the microphones indicated in the figure.

Recording of the sound with two separate microphones

Let us now return for a moment to the experimental setup of fig. 2. Instead of recording the sound in this case with two microphones mounted on an

³⁾ See the article: Stereophonic sound reproduction, Philips techn. Rev. 5, 107, 1940.

artificial head, it is also possible to use two separate microphones. Practically the same stereophonic effects are obtained, provided the separate microphones are placed about three times as far apart as is the case with the microphones on the artificial head. In a manner analogous to the one used above, the following rule of thumb is found: (distance between separate microphones) $\times$ (angle to be projected in degrees) = 6000, *i.e.* with a constant three times as large as in the former case. See *fig. 8*.

The fact, that the constant must be larger, is easily understood, when it is kept in mind, that two separate microphones at the same distance apart as the microphones of an artificial head would register a considerably smaller ratio of intensities and would thus give considerably less stereophonic effect. In the case of the artificial head the ratio of the intensities at the position of the microphones is considerably increased by reflection and diffraction.

Although the separate microphones have the advantage, that their mutual distance can be changed quickly upon passing from one object to be projected to another, while otherwise a different artificial head must be chosen each time, the artificial head is, nevertheless, preferable, because it is found to give sharper sound images. This is in agreement with the experimentally found result, that a sound image is sharper, the more it is determined by the intensity ratio rather than the time difference. With the separate microphones which must stand farther apart, time differences play a relatively large part and the result is somewhat less sharp images.

The characteristics here mentioned have been investigated with microphones whose sensitivity did not depend upon the direction from which the sound comes. The peculiarities, which occur when

the microphones possess a certain directional effect, will not be considered, nor will the generally slight effect of the dimensions of the microphones.

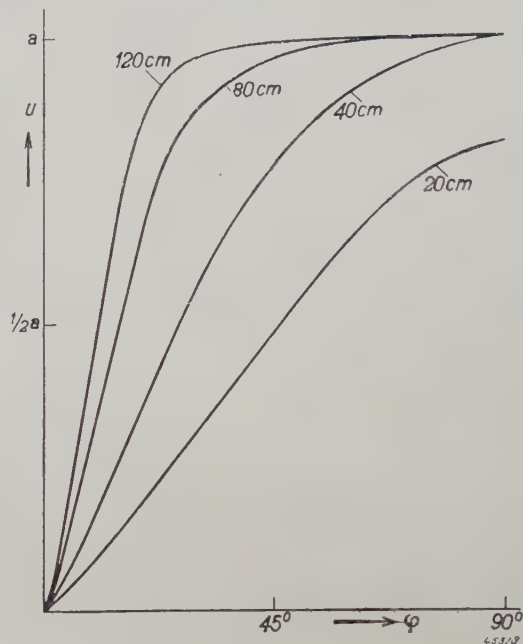


Fig. 8. In recording the sound of a source *G* (*fig. 2*), which is placed at an angle φ , with regard to the plane of symmetry, *u* represents the direction of the sound image from the centre observed by *W* in the reproduction room. The different curves belong to the distances recorded in the figure between the two microphones.

In conclusion, we should like to point out, that in several very special cases the reproduction of the sound may be obtained, not by means of two loudspeakers, but by means of two headphones, one for the right and one for the left ear. For this case, which until now, has only been applied in the case of persons with poor hearing, analogous rules for the projection can be set up. We shall not go into it here, since this method of reproduction has a much less general application.

SEVERAL AFTER-EFFECT PHENOMENA AND RELATED LOSSES IN ALTERNATING FIELDS

by J. L. SNOEK and F. K. du PRÉ.

537.226.31

The combatting of the dielectric and magnetic after-effect is often an important problem, especially in the development of new materials for electrotechnology. After-effect phenomena are the cause of the „losses” in dielectrics, while at least part of the losses in ferromagnetic substances must also be ascribed to that cause. In this paper the phenomena are described, which occur in the after-effect and their causes are discussed, while at the same time the close analogy and the connection are pointed out between electrical and magnetic after-effect and the phenomena of elastic after-effect.

The search for materials, which possess new properties or new combinations of already known properties, which often make possible entirely new constructions, is becoming more and more important. In electrotechnology this search is manifested for example in the extensive investigation in the field of dielectrics and ferromagnetic substances. However, no matter what new properties it may be desirable for these electrotechnical materials to possess, one fundamental requirement will always be made, namely, that they must show very small „losses”.

This requirement is understandable, when it is kept in mind, that in a dielectric or ferromagnetic substance, which is not free of losses, a certain heat development will occur, when it is placed in an alternating electric or magnetic field. This heat development may, in the case of strong alternating fields, such as may for example occur in transmitters, mean a considerable loss of energy, while at the same time the material may be damaged by the too rapid rise in temperature. Also in those cases, where the energy loss itself is no objection, the heat development often has a harmful effect, it is for example known, that the very important selectivity of oscillation circuits or filters is diminished by the use of materials, which are not free of losses. It is for this reason, that the combatting of losses or the preparation of new substances with few losses is such an important technical problem.

The nature of the losses in dielectrics and ferromagnetic substances may be varied. In a dielectric, when the direct-current conductivity of the substance may be neglected, the losses must practically always be ascribed to after-effect phenomena.

In a ferromagnetic substance, on the other hand, the losses are caused partly by the conductivity of the substance (eddy-current losses), while hysteresis may be mentioned as a second important cause.

Finally in these substances also a part of the losses must be ascribed to after-effect, although this part is often relatively small. In the development of new magnetic materials, however, it has repeatedly been found, that a substance with otherwise favourable properties could not be employed, because the strong after-effect phenomena occurring in it could not be combatted effectively. To mention a single example, the transformer iron, prepared electrolytically, is still very little used in heavy-current technology, only because of the disturbing after-effect, which is difficult to combat.

The investigation of the after-effect phenomena is therefore important not only in the case of dielectrics, but also in that of magnetic substances.

It is now our intention to show, that matter may exhibit phenomena of after-effect, not only under the influence of electrical or magnetic forces, but also under the influence of mechanical forces, demonstrating at the same time, that the after-effects are caused in many cases by a diffusion phenomenon.

Elastic after-effect

As an example, we have chosen a case of elastic after-effect, because the phenomena here are easy to explain and at the same time show a close analogy with other after-effect phenomena.

The following experiment is performed: A rod clamped at one end is suddenly bent by a force acting at the other end. It would be expected, that when the bending force is kept constant, the bending would also remain constant. This is not, however, the case; the bending gradually increases somewhat, to reach a new true equilibrium value: elastic after-effect occurs¹⁾ (*fig. 1*).

This elastic after-effect is very general and in different materials, it is often due to the same cause, namely, heat conduction. If we assume, that the material has a positive coefficient of expansion,

the side, which is suddenly compressed will, as we learn from thermodynamics, become slightly warmer, and the other side slightly colder. The compressed, warm side will then exhibit extra resistance

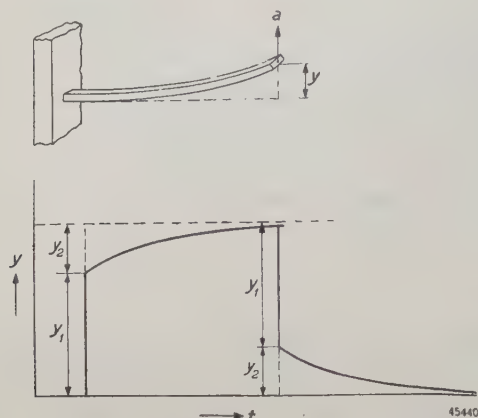


Fig. 1. At the end of the above sketched rod a constant force a is suddenly applied, and after the elapse of a long time suddenly removed again. The course of the bending y as a function of the time is indicated in the lower half of the figure. The after-effect is very much exaggerated, ordinarily y_2 amounts to less than 1% of y_1 .

to further compression, because it has a tendency to expand due to the rise in temperature. Similar reasoning is valid for the side that is stretched, which has a tendency to shrink. As soon as the difference in temperature is equalized by the conduction of heat, the elastic force decreases slightly and the rod bends farther. The phenomenon is known as the thermo-elastic effect.

In the exceptional case of a negative coefficient of expansion the warm side is the stretched side, but otherwise the reasoning is the same.

Closer analysis shows, that the equalization of temperature takes place in such a way, that the change in temperature ΔT at every point in the rod decreases as $\Delta T e^{-t/\tau}$. In this connection the bending y due to a constant force a can be represented by:

$$y = y_1 + y_2 (1 - e^{-t/\tau}). \quad (1)$$

This equation expresses the fact, that at $t = 0$ a bending I_1 occurs, while after some time the bending has increased to $y_1 + y_2$. The extra bending y_2 occurs gradually, corresponding with the change in temperature. The quotient y_2/y_1 which is usu-

ally very small (for instance 1%), may be called the strength of the effect. The quantity τ is called the relaxation time.

Upon the sudden removal of the external force, the deviation at first suddenly decreases by an amount y_1 , while the additional y_2 gradually disappears according to $y_2 e^{-t/\tau}$. From this, we conclude, that the substance suffers no permanent alteration due to the after-effect. On the contrary, the original condition is completely restored.

When the force a acting on the end of the rod is no longer constant, but periodical:

$$x = a \cos \omega t, \quad (2)$$

the deviation can be represented by:

$$y = b \cos (\omega t - \delta), \quad (3)$$

as long as $\omega \ll$ the characteristic frequency of the rod. In the case of a periodical force, therefore, the result of the after-effect is, that a phase difference occurs between force and deviation. This important property can be derived in a simple way from the above facts, as will be shown below.

The following holds for the phase shift:

$$\tan \delta = \frac{y_2}{y_1} \frac{\tau \omega}{1 + \tau^2 \omega^2}, \quad (4)$$

while the amplitude b lies between y_1 and $y_1 + y_2$, as follows from:

$$b = y_1 + \frac{y_2}{1 + \tau^2 \omega^2} \quad (5)$$

The quantities y_1 and y_2 are the same as those appearing in formula (1) for the case of a constant force a .

The value of δ reaches a maximum for that frequency for which $\tau \omega = 1$, see fig. 2. Expression (4) is valid only for small values of y_2/y_1 , which is almost always the case in practice.

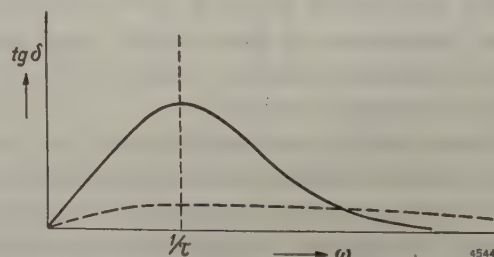


Fig. 2. With periodic loading of the rod shown in fig. 1 the likewise periodic deviation exhibits a phase shift δ (retardation), with respect to the force. The quantity $\tan \delta$ is here plotted as a function of the angular frequency ω . The full-line curve holds for the simplest case of after-effect, the broken line curve for a more complicated case. τ is the relaxation time.

¹⁾ In this experiment in order to prevent the rod from vibrating about its equilibrium position, thus to avoid effects of inertia, several precautions must be taken. The characteristic time of oscillation of the rod must be very short compared with the time during which a noticeable after-effect occurs. In that case the „sudden” force can be applied in a time which is very short compared with the after-effect time, but at the same time very long compared with the characteristic oscillation time of the rod, so that the after-effect is not disturbed by the occurrence of characteristic vibrations.

It must be noted that, due to the after-effect, in this periodical movement heat is developed in the rod. The work performed by the force x per period, $\oint x dy$, is, due to the phase shift, not equal to zero. The after-effect causes a loss of energy. Furthermore it may be seen from fig. 2, how the influence of the after-effect can be combated, not only by removing the cause, but also by employing very high or very low frequencies.

If the after-effect does not have the simple form indicated by (1) and fig. 1, the after-effect curve can, nevertheless, always be analyzed and expressed in the form:

$$y_1 + \sum_n y_n (1 - e^{-t/\tau_n}) \dots \dots \dots (6)$$

We then again find for the deviation under the influence of a periodical force:

$$y = b \cos (\omega t - \delta), \dots \dots \dots (7a)$$

where now, however,

$$\operatorname{tg} \delta = \frac{\sum_n y_n}{y_1} \frac{\tau_n^2 \omega}{1 + \tau_n^2 \omega^2} \dots \dots \dots (7b)$$

and

$$b = y_1 + \sum_n \frac{y_n}{1 + \tau_n^2 \omega^2} \dots \dots \dots (7c)$$

The variation of $\tan \delta$ with the frequency may be much flatter than in the case of a single relaxation time, see for instance the broken line in fig. 2.

We shall now briefly derive the above results. For the sake of simplicity we assume, that the after-effect curve can be described by a single power of e :

$$y = y_1 + y_2 (1 - e^{-t/\tau}).$$

Since y_1 as well as y_2 will be proportional to the constant force a , we may also write:

$$y = a [\varepsilon_1 + \varepsilon_2 (1 - e^{-t/\tau})],$$

or more briefly:

$$y = a [\varepsilon_1 + \varepsilon_2 f(t)].$$

In order to derive from this the fact, that in the case of periodical loading the above-mentioned phase shift occurs, we must still call attention to an experimental peculiarity, namely the superposition effect, which is found not only here, but in all other after effect phenomena in solids.

Let us assume, that a suddenly applied loading is altered after a certain time t_0 by a different suddenly applied loading (for example in the opposite direction). For the deviation now occurring we find, as sketched in fig. 3, a form, which is given by the superposition of the deviations, which each of the processes by itself would have caused. If the first deviation would have been represented at a given moment t by $\varepsilon_1 a + \varepsilon_2 f(t) \cdot a$, and the second by itself would have been

$\varepsilon_{10} + \varepsilon_2 f(t - t_0)b$, then the final result can be found by adding these expressions.

If the loading is given not by „impulses”, but by a continuous function $x(t)$, then an increase in loading dx , occurring

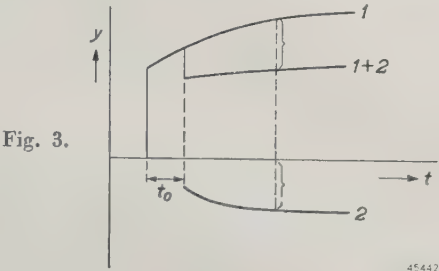


Fig. 3.

at the moment t , will be manifested at a later moment t_1 by a contribution to the deviation of the amount:

$$\varepsilon_1 dx + \varepsilon_2 f(t_1 - t) dx.$$

In order to deal with the above-described case, we now assume that from the moment, $t = \text{„zero”}$ the following will, hold: $x = a \cos \omega t$. As fig. 4 shows, at $t = \text{„zero”}$ we must begin by introducing a jump of magnitude a in the loading. According to the superposition principle the deviation at any given moment t_1 is then given by:

- 1) The contribution of the jump, thus $\varepsilon_1 a + \varepsilon_2 f(t_1)a$;
- 2) The contribution of the continuously varying loading:

$$\varepsilon_1 \int_0^{t_1} dx + \varepsilon_2 \int_0^{t_1} f(t_1 - t) dx, \text{ with } x = a \cos \omega t.$$

These contributions can be calculated in a simple way. We find:

$$y = a \left(\varepsilon_1 + \frac{\varepsilon_2}{1 + \tau^2 \omega^2} \right) \cos \omega t_1 + a \varepsilon_2 \frac{\tau \omega}{1 + \tau^2 \omega^2} \sin \omega t_1 - \frac{a \varepsilon_2}{1 + \tau^2 \omega^2} e^{-t_1/\tau}.$$

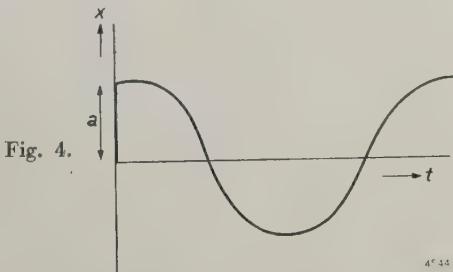


Fig. 4.

The last term may always be neglected in practice, while the other two can easily be reduced to the form already discussed, where the arbitrary moment t_1 is again called t . In the case of more complicated after-effect according to (6) the derivation is analogous.

In connection with a case of magnetic after-effect to be discussed later, we should like to consider very briefly one of the many other phenomena of elastic after-effect, for which purpose the following experiment is performed.

An iron wire is suddenly given a torque by means of a constant couple. The angle of torsion shows a variation like that represented in (1),

thus first a sudden increase and then a gradual further increase²⁾.

Although the phenomenon proceeds in a manner entirely analogous to the previous one, the cause must certainly not be sought in the thermo-elastic effect, because no change in volume and thus no temperature differences occur upon torsion. The cause only became clear, when it was found, that the phenomenon was only observed in the case of iron, which contains small quantities (0,01%) of carbon or nitrogen in solid solution. The "dissolved" particles are situated at intermediate lattice spots. The way, in which these dissolved particles can cause the after-effect, can be explained as follows.

Upon torsion each volume element of the iron experiences a gliding, which amounts to a compression of the lattice in one direction and an expansion in the perpendicular direction. The carbon or nitrogen particles, situated between the iron atoms, therefore diffuse from the compressed spots, where little space is left, to the stretched spots. Due to this, however, a decrease in the elastic tensions occurs at the same time, so that a constant couple is able to give the wire gradually a further torque, corresponding to the progress of the diffusion.

In the two examples, mentioned here, the cause of the after-effect was a diffusion phenomenon, in the first a diffusion of heat, which restored a disturbance of the uniform temperature, in the second a diffusion of material particles, which compensated for a disturbance of the concentration. This property is not confined to the examples given. On the contrary, in almost all cases of after-effect in solids investigated until now, it has been found that the after-effect is based on a diffusion process, which restores a disturbed thermodynamic equilibrium.

The phenomena are always quite analogous to the ones described above, as will appear later in the case of the electrical and magnetic after-effect, to which we shall now devote our attention.

Incidentally, it may be noted, that the establishment of a thermodynamic equilibrium will always take a certain time. After-effect will only be observed experimentally, when the equilibrium is established wholly or in part at a sufficiently slow rate (e.g. in a time comparable with the characteristic time of the considered experiment).

Dielectric after-effect

The experiments, in which the dielectric after-effect was originally first observed many years ago, are entirely analogous to those used to demonstrate the elastic after-effect. We consider a dielectric, which is suddenly placed in an electric field. The dielectric displacement D which thereupon occurs, exhibits in almost all cases a variation analogous to that of (1), i.e. first a sudden and then a gradual increase. This, often very rapid, phenomenon is shown diagrammatically in fig. 5, as well as the (less simple) variation of the somewhat more graphic quantity $I = 1/4\pi \cdot dD/dt$,

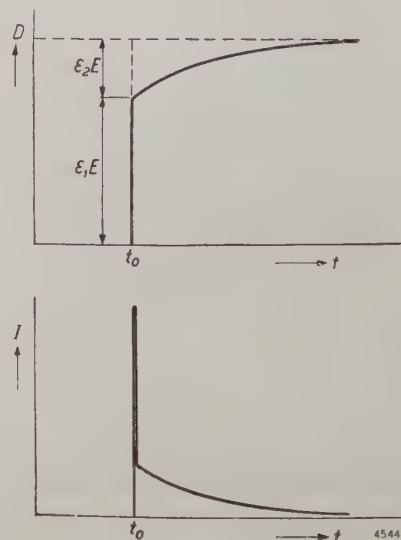


Fig. 5. Variation of the dielectric displacement D and of the displacement current $I = 1/4\pi \cdot dD/dt$ in an insulator, which at the moment $t = t_0$ is suddenly placed in an electric field.

the displacement current in the dielectric. The behaviour of D upon the sudden disappearance of the field will be immediately clear. Although the variation of the after-effect is often not according to a single power of e , but one which must be described by (6), we shall first confine ourselves to the simple case.

From the observed after-effect we conclude, as above, that in a dielectric placed in a periodical field:

$$E = E_0 \cos \omega t \dots \dots \dots (8)$$

a dielectric displacement takes place according to:

$$D = D_0 \cos (\omega t - \delta) \dots \dots \dots (9)$$

with:

$$D_0 = E_0 \left[\epsilon_1 + \frac{\epsilon_2}{1 + \tau^2 \omega^2} \right] \text{ and } \tan \delta = \frac{\epsilon_2}{\epsilon_1} \frac{\tau \omega}{1 + \tau^2 \omega^2},$$

where ϵ_1 and ϵ_2 are constants.

²⁾ This phenomenon was first observed by Becker, and Kornetsky in carbonyl iron; every other kind of iron, however, shows the same behaviour.

The fact, that the dielectric displacement D can be expressed, in the manner here given, is the most important practical result of its interpretation by means of after-effect.

The following is valid for the current density:

$$I = \frac{1}{4\pi} \frac{dD}{dt} = -\frac{D_0}{4\pi} \omega \sin(\omega t - \delta) = \frac{D_0}{4\pi} \omega \cos\{\omega t + (90^\circ - \delta)\}.$$

Since the current is not exactly 90° in phase ahead of the voltage, but only $90^\circ - \delta$, heat development occurs in the substance, the amount being determined by the magnitude of δ . When we express graphically the relation between D and E according to (8) and (9) we obtain a slender ellipse as drawn in fig. 6. Its width is determined by δ , while its area indicates the energy development per period.

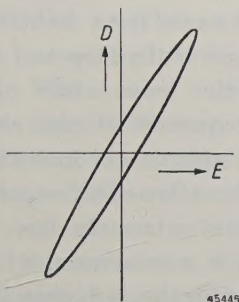


Fig. 6. Relation between the dielectric displacement D and the field strength E for a dielectric, which is placed in a periodic electric field and which is not free of losses.

When we are not interested in the energy development alone, it is often clearer to describe the influence of the angle of loss in a different way.

It must be noted, that the same phase difference $90^\circ - \delta$ would have occurred between current and voltage, if a condenser with a loss-free dielectric were connected in series or in parallel with a certain resistance. This apparent series or parallel resistance, which presents difficulties in various electrical connections, is a graphic measure of the harmful effect of the losses.

Attempts have been made, for more than fifty years, to make the phase shift δ as small as possible. The requirements, which must be made, are often very severe. Thus a value of $\tan \delta$ of 0.01 must already be considered large, although this corresponds to a phase angle of less than one degree. In the best solid dielectrics, now known, a value of δ of about 0.5 minute has been attained ($\tan \delta \approx 10^{-4}$).

It must, however, be noted that just in these technically much used substances the after-effect

must often be described by expressions analogous to (7), thus by:

$$D = D_0 \cos(\omega t - \delta),$$

with:

$$D_0 = E_0 \left[\epsilon_1 + \sum_n \frac{\epsilon_n}{1 + \tau_n^2 \omega^2} \right] \text{ and}$$

$$\tan \delta = \sum_n \frac{\epsilon_n}{\epsilon_1} \frac{\tau_n \omega}{1 + \tau_n^2 \omega^2}.$$

The variation of $\tan \delta$ with the frequency is then often, such as that indicated by the broken line in figure 2, in other words, $\tan \delta$ depends little on the frequency.

Cause of the dielectric after-effect

It is assumed, that the after-effect losses are generally caused by „imperfections”, namely either by regions in the substance which possess a certain (slight) electrical conductivity, or by dipoles, present in low concentration in the substance, which can be oriented by an electric field with a certain time lag.

It is not difficult to understand, that a substance with these properties would exhibit just the phenomena observed. A suddenly applied electric field first causes spontaneously the „ordinary” dielectric displacement E , while afterwards the quantity $D (= E + 4\pi P)$ gradually increases farther, either due to the orientation of the dipoles, which increases the polarisation P , or due to the fact, that in the above-mentioned regions the positive charge accumulates on one side and the negative on the other, likewise resulting in an increase of P .

A medium, in which all the dipoles or all the regions possess equal properties, will then exhibit a simple after-effect curve according to (1), while the complicated case occurs, when the dipoles are mutually different as regards their time of adjustment, or the regions in their conductivity.

In both case, the after-effect is based upon processes, which are closely related to diffusion. Because when a dipole, consisting as it does of a positive and a negative part, changes its orientation, at least one of the two parts must change its place in the lattice, which displacement takes place only gradually under the influence of the electric field and thermal agitation, as is characteristic of diffusion. In the other case we are concerned with a diffusion of ions or „diffusion” of electrons.

Ferromagnetic after-effect

While the existence of the dielectric after-effect is quite generally known, the same can certainly

not be said of the magnetic after-effect, which is, nevertheless, also of great importance in alternating current technology. In magnetic materials, aside from the eddy currents, the more obvious phenomenon of hysteresis has drawn the main attention.

In connection with the development of new materials, however, it is important to pay especial attention to the after-effect.

Because of the fact that, in a ferromagnetic substance all three of the effects mentioned usually occur at the same time, the first important task is to be able to separate these effects from each other. Fortunately in this respect, with a suitable form of the ferromagnetic substance, the influence of the eddy currents on the phenomena in an alternating current circuit can be determined theoretically. This influence can thus be separated from the other two effects. Moreover, because this influence can always be kept small by employing the substance in lamellar form, we shall in the future disregard the eddy currents.

We shall now compare the two other effects, hysteresis and after-effect with each other, which comparison will furnish us with a means of determining the influence of the two phenomena separately. We assume, that the substance is placed in an alternating magnetic field. The hysteresis characterized by the fact is that the phase of the magnetic induction B is behind that of the field H . The relation between B and H being given by the hysteresis loop (fig. 7a).

In hysteresis, as well as in the after-effect, energy is developed in the substance, and this is given by the area of the hysteresis loop or of the ellipse, respectively. There are thus various points of correspondence.

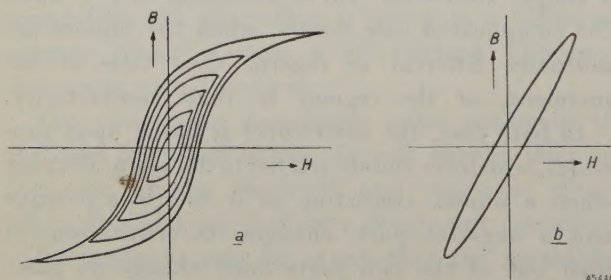


Fig. 7. a) Hysteresis curves of a ferromagnetic material. The variation B is drawn, when the field is varied from $+H$ to $-H$ and back again. The different curves are for different amplitudes of the variations of the magnetic field b) Relation between the magnetic induction B and the field strength H for a ferromagnetic substance placed in an alternating magnetic field and which exhibits after-effect only. Usually hysteresis and after-effect occur together, which makes the phenomena less clear.

An important difference between the two effects occurs, however, when we make the amplitude of

the field smaller. The ellipse, as well as the loop, then naturally become smaller. The ellipse, however, retains its original proportions. The hysteresis loop, on the other hand, is deformed and approaches a straight line; its area thus decreases much more rapidly than of the ellipse. With very small amplitudes of the field, we thus retain only the influence of the after-effect, which consists in a phase shift and consequent heat development. With the help of this property we can separate the two effects.

Another phenomenon may also occur when the frequency of the field, instead of its amplitude, is changed. The after-effect ellipse may then become broader or narrower, as in the dielectric case, because, at least in certain frequency regions, the size of the phase angle depends upon the frequency. The hysteresis loop in this case remains unchanged.

Summarizing, we may therefore say, that the relation between B and H in the case of hysteresis and after-effect exhibits a difference, firstly in the shape of the loop and then also in the way in which the loops alter upon change in amplitude or frequency of the alternating field.

In order to be able now to investigate the phenomena of magnetic after-effect separately, we shall always work with extremely low field strengths, where hysteresis is a minimum, while a correction will be introduced for the influence of eddy currents.

Suppose now, that we perform the following experiment. The substance is suddenly placed in a weak magnetic field. It is then found, that the induction B shows a variations with time like that, which we have already encountered several times, and which can be represented by a curve like that of fig. 5, i.e. a sudden increase followed by a gradual increase. (The variation of dB/dt , which determines the induction voltage in neighbouring conductors, is entirely analogous to that of dD/dt , which is also drawn in the figure).

From this we again conclude in the familiar way that with a periodical variation of the field:

$$H = H_0 \cos \omega t, \dots \dots (10)$$

the following is valid for the induction:

$$B = B_0 \cos (\omega t - \delta) \dots \dots (11)$$

The validity of this expression, for B , is the most important practical result of the after-effect. The relation between B and H , as may be seen from (10) and (11) is given by the ellipse of fig. 7b, as already mentioned. The occurrence of the phase shift δ results in the fact, that in a coil in which the substance in question is employed the alternating

current will no longer be 90° in phase behind the voltage, but slightly less. Just as in the dielectric case, it may be said here also that the same phase difference would occur if the ferromagnetic substance were free of losses and the coil connected with a certain resistance in series or with a certain resistance in parallel. This apparent series or parallel resistance, which in practice is often objectionable, is a measure of the harmful effects of the losses.

Cause of the ferromagnetic after-effect

Before we discuss the cause of the after-effect for a single thoroughly investigated case, we shall first briefly recall the remarkable nature of the structure of a ferromagnetic substance with not too large a coercive force.

Such a substance is built up of little regions each of which is completely magnetized, but whose directions of magnetization are very different. In a single crystallite these directions are at angles of 90° and 180° with each other, since the magnetization occurs according to the cube directions in the crystal. Where two such regions, whose dimensions are about 10^{-4} cm, border on each other, a gradual transition occurs from the magnetization direction of the one to that of the other region. If the substance is now placed in a weak magnetic field, this transition region, the „wall”, is slightly displaced in such a way, that the region whose direction of magnetization corresponds more closely to the field becomes slightly larger and the other slightly smaller, whereby the substance becomes magnetized. The direction of magnetization is thus altered at certain spots and this causes a unilateral deformation of the crystal as has been demonstrated by experiments on magnetostriction. The elastic forces tend to prevent this deformation, which stretches the lattice in one direction and compresses it in the perpendicular direction. It is therefore, remarkably enough, the elastic forces which limit the extent of the magnetization brought about; the way in which this is done is, however, not known with certainty.

When the field is increased to such a value, that the wall would be displaced over a certain critical distance, the wall suddenly shoots forward and the magnetization in the remainder of the „defective” region as a whole changes over to the direction of the „better” region. This effect lies at the bottom of hysteresis. In connection with the following it must be noted, that a change of 90° or 180° in magnetization causes no change in the mechanical tensions in a given crystallite. The changes in tension upon magnetization thus do not appear

in the regions, but only in the transitional layers, so that only hysteresis accompanies the „change-over” in direction.

After this outline of the structure of ferromagnetic substances, we return to the causes of the magnetic after-effect. The most thoroughly investigated case is that of the above-mentioned iron wire, where the following conception has been arrived at.

When this substance is suddenly placed in a weak field, or more generally, when the field H_0 in which the substance is situated is suddenly increased by h , the „walls” are suddenly shifted, and therefore local deformations in the lattice suddenly occur: a stretch in one direction, a compression in the other. The carbon and nitrogen particles situated between the iron atoms therefore diffuse from the compressed spots, where there is little room left, to the stretched spots. This, however, causes at the same time a decrease in the elastic tensions and the field can then gradually displace the wall somewhat farther, increasing the magnetization thereby, which is exactly what is observed in the after-effect. The remarkable feature of the description here given of the magnetic after-effect³⁾ is that the actual cause lies in the elastic after-effect of the material.

It must be noted that here also it is not always the simple relation (4) which is observed for $\tan \delta$, but often an expression like (7) occurs, with a much flatter shape as a function of the frequency. In this case also, perhaps the cause must be sought in analogous diffusion phenomena.

Thus as we have just seen, it is not necessary that the after-effect in magnetic materials should be based on „purely magnetic” effects; usually indeed this is not the case. If there is any coupling between the effect being investigated and another which exhibits after-effect, after-effect phenomena will also appear in the first effect. This situation is often encountered in the magnetic case.

Another example is the following. When a so-called powder core, i.e. an aggregate of magnetic particles separated from each other by a thin insulating layer, is magnetized, elastic forces will in general appear as a result in the insulating layer. Now, if the insulating layer exhibits elastic after-effect, this will again be manifested in the occurrence of a phase difference in the magnetic phenomenon.

³⁾ There exists, however, still another magnetic after-effect, that of Jordan, for which this does not hold. We merely mention this fact, without going into any details.

It is understandable that the detection of the causes of losses is often made very difficult by such couplings. In the meantime the examples mentioned may suffice and we shall not discuss any other examples.

Our final conclusion is, that the material may exhibit after-effect under the influence of mechanical as well as of electrical or magnetic forces, while the cause must often be sought in a diffusion

process. In the case of periodical electrical or magnetic fields the most typical manifestation of after-effect is the phase difference exhibited by the dielectric displacement D or the magnetic induction B with respect to the field, and the losses connected with this phase difference. The diminution of this phase difference is therefore of great importance in electrotechnology, and this has already been successfully accomplished in various cases.

Contents of *Philips Research Reports* 1, No. 2. January 1946

- C. J. Bouwkamp: Radiation resistance of an antenna with arbitrary current distribution.
- M. J. Druyvesteyn: Some remarks on the thermal expansion and the Poisson contraction of the cubic metals.
- P. H. J. A. Kleynen: Extension of Langmuir's (ϵ , η) tables for a plane diode with a maxwellian distribution of the electrons.
- A. van der Ziel: Extension and application of Langmuir's calculations on a plane diode with maxwellian velocity distribution of the electrons.
- F. M. Penning and J. H. A. Moubis: The contraction phenomenon in a neon glow discharge with molybdenum cathode.
- J. M. Stevels: Scale glass as a substitute for mica.
- C. J. Bouwkamp and N. G. de Bruijn: The problem of optimum antenna current distribution.